Molecular and Genetic Characterisation of a Small Segment of the *Drosophila melanogaster* X chromosome

By

Ute Christiane Schuppler

A thesis submitted for the degree
of Doctor of Philosophy of the
Australian National University

February 1992

NO PAGE 33 TEXT OK.

Widmung

für

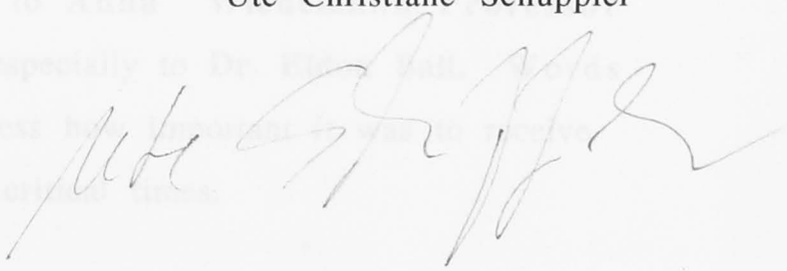
Margrit und Klaus

ohne die es diese Arbeit
nicht gegeben hätte

Declaration

In accordance with the regulations of the Australian National University, I wish to state that the work described in this thesis was carried out by myself except where due reference is made.

Ute Christiane Schuppler



Acknowledgments

It is a pleasure for me to express my sincerest thanks to **Dr. John Gibson, Dr. Steven Delaney** and **Dr. David Hayward** who showed me how stimulating and enjoyable science can be. I am very grateful for their encouragement, patience and support. But, I am especially appreciative for their interest in my work and the ensuing exciting discussions.

Many thanks to **Dr. David Hayward** who introduced me to the nitty gritty of RNA work, **Dr. Steven Delaney** who encouraged me to pursue the embryo injections and **Jane Olsen** who held my hand in all sequencing matters.

I am deeply indebted to **Anna Wiedemann, Professor Barry Osmond** and especially to **Dr. Eldon Ball.** Words are insufficient to express how important it was to receive their moral support in critical times.

Finally, I would like to thank all of my friends for putting up with me during the last critical weeks and even more for sharing the fun parts throughout the last four years. We had a hell of a good time.

I acknowledge the contribution made by Dr Miklos in suggesting the project and in the provision of essential materials and laboratory facilities.

Abstract

Two adjacent *Drosophila melanogaster* complementation groups, *A112* and *LB20*, were located within a chromosomal walk in the region of polytene band 19F/20A. Mutations at both genes are recessive lethals (Lifschytz and Falk, 1968; Schalet and Lefevre, 1976). All alleles described at the *A112* locus were found to be embryonic lethals and those at the *LB20* locus were lethal in the late larval stages. Although the region itself is well characterized genetically, little was known about these two lethal genes.

The molecular positions of *LB20* and *A112* were determined using deletion mapping. This analysis located the gene *LB20* to a 53kb DNA region between the breakpoints of *Df(1)HM44* and *Df(1)JC4*. As it was known that heterozygotes *LB20/Df(1)Q539* and *LB20/Df(1)JA117* were lethal, the location of *LB20* could be further defined to a 15kb DNA region. Similar analyses located *A112* between the proximal breakpoint of *Df(1)16-129* and the distal breakpoint of *Df(1)JC4*. Using data on transcripts produced by genes known to be adjacent to *A112*, the locus was shown to be located in the region of overlap between the two deficiencies, *Df(1)JC77* and *Df(1)HM44*.

Transcription analyses revealed two transcripts, one of 2.2kb the other of 0.6kb, within the *A112/LB20* region. cDNAs corresponding to both transcripts were isolated and sequenced. A cDNA of 1.95kb corresponding to the sequence of the 2.2kb transcript was sequenced. Comparison of the predicted amino acid

sequence to known proteins revealed that the putative A112 protein contained the sequence motifs which characterize the RNA helicase family.

A 0.6kb transcript discovered in the *A112/LB20* region was initially postulated to encode the *LB20* gene. It was known that this transcript contained sequences which hybridized to a *Hind* III fragment identified with a probe corresponding to a chicken collagen gene. cDNA and genomic DNA corresponding to this "collagen-like" gene were isolated and sequenced. Analysis of the nucleotide sequences and the predicted amino acid sequences showed that the similarity to the collagen gene was superficial and based solely on a sequence repeat (GGX GGX CCX). The protein encoded had not previously been described in the data banks and its function remains to be elucidated.

Two constructs, each containing 12kb from the *A112/LB20* region were introduced via P-elements into the germline of *Drosophila* embryos and transgenic lines were generated. Transgenic males produced from the first construct, which contained the genomic DNA corresponding to the 2.2kb transcript and a 5kb 5' region, were crossed to virgin females carrying an *A112* allele. From these crosses the male offspring carrying the *A112* mutant allele were viable, indicating that the normal function of the *A112* locus had been restored by the inserted DNA.

Transgenic lines were obtained from the second construct which included DNA corresponding to the 0.6kb transcript together with 9kb from the 3' region. The DNA in this second construct overlapped that in the first construct by 6kb. It was found that

both constructs contained the DNA necessary to restore the normal function at the *LB20* locus. Thus the *LB20* gene was localized within the 6kb overlap between these two constructs. A preliminary experiment using poly(A)$^+$RNA provided evidence for a 3kb transcript in the region of overlap between the two transformation constructs. It is possible that this transcript is derived from the *LB20* gene.

Contents

Chapter 1

General Introduction

1. General Introduction

The number of genes in an organism is one of its fundamental biological parameters and relates to the number of functions required to construct that organism. It has generally been assumed that with increasing organic complexity the number of genes of the species increases. The possibility does exist however, that increasing complexity results from the effects of interactions between a small and relatively constant number of genes. It is for questions like this that knowledge of the number of genes in any organism is regarded as important.

In *Drosophila melanogaster* the question of the number of genes was closely connected to the number of polytene bands. Early studies on the morphology of the polytene chromosomes of *Drosophila melanogaster* and their characteristic banding pattern led to the hypothesis that each of the individual polytene chromosome bands is associated with a single gene locus with the implication that the number of genes could be determined simply by counting the number of bands (Painter, 1933; Bridges, 1935; 1938). Many attempts were initiated to validate or discredit this "one-band one-gene" hypothesis (Gausz *et al.*, 1979, 1986; Woodruff and Ashburner 1979; Hilliker *et al.*, 1980; Lefevre 1974). The calculations used to analyse this hypothesis were based on inducing lethal mutations (Muller, 1927; Auerbach 1947). The goal was to determine the number of mutable loci within a small segment and then compare this number to the number of bands in that region of the map. The debate on this point continued for 50

years until conclusive evidence was presented that the number of genes was not simply related to the number of bands (Lefevre and Watkins, 1986).

The vast majority of the lethal genes identified in these screens were recessive. In those cases where it was not possible to associate lethal alleles with visible phenotypes it was very difficult to analyse the genes further, apart from determining the stage of lethality and the number of lethal complementation groups. With the advent of molecular biological techniques the results from a number of screens of lethal genes have been combined with analyses of the mutants at the molecular level, using as a basis the substantial body of knowledge obtained from genetic mutagenesis studies.

The molecular analyses of lethal complementation groups in general followed two approaches. One, was to focus on the molecular organisation of a specific region of the chromosome, e.g. the study of the *rosy-Ace* region. Instead of comparing the number of lethal complementation groups to the number of bands, the molecular approach focused on the number of transcripts as molecular units. The second approach analysed lethal alleles in terms of the period during which they were lethal.

One of the first examples of the first approach was the analysis of division 87 on the third chromosome by Gausz *et al.* (1979). Within this division maps the *rosy-Ace* region, named after *rosy*, the gene for xanthine dehydrogenase (Chovnick *et al.*, 1977), and *Ace*, the gene for acetylcholinesterase (Hall and Kankel, 1976). In a study which attempted to saturate the region with lethals alleles, it was found that there were 21 complementation

groups in an interval containing 23-26 bands (Hilliker *et al.*, 1980). This was initially confirmed when 315kb of DNA in the region was cloned and 12 complementation groups were located which corresponded to 13 of Bridges' bands (Bender *et al.* 1983; Spierer *et al.* 1983). However, the following analysis of transcriptional activity in the cloned region identified a total of 43 transcription units, over three times more than the number of bands (Hall *et al.*, 1983; Bossy *et al.*, 1984).

In other sections of the genome similar detailed studies, aimed at the analysis of chromosomal structure and the chromosomal distribution of functional genes, also utilised a combination of classical genetic and molecular approaches. Examples are studies of subdivision 2C to 3C (Shannon *et al.*, 1974; Perrimon, *et al.*, 1984b, 1985); division 19 and 20 (Schalet and Lefevre, 1976; Lefevre and Watkins 1986; Perrimon, *et al.*, 1989) division 36 (Steward and Nusslein-Volhard, 1986), and subdivision 84B-C (Lewis *et al.*, 1980; Cavener 1986).

Equally successful was the second approach in which a systematic search was conducted for mutations lethal at specific developmental stages. For example a systematic search for embryonic lethals alleles identified 15 loci, each affecting the number of segments (Nusslein-Volhard and Wieschaus, 1980). These loci were the first to be described in which pattern formation was affected in a specific way. They are divided into three classes: the gap mutants, the pair-rule mutants and the segment polarity mutants. These classes of genes build, together with the maternal genes, a regulatory cascade which governs pattern formation in the developing embryo (Nusslein-Volhard and Wieschaus, 1980; Ingham, 1988). As the intricate relationship of

these genes with the genes of the *bi thorax* and *Antennapedia* complexes was discovered, a model emerged that provided details of the molecular processes fundamental to cell specification. This model also describes the interactions of genes, their products and the hierarchical order of events leading from a protein gradient in the egg to segment formation in the embryo (Pfeifer *et al.*, 1987; Ingham, 1988).

A second example is the work of Gatti who found that mutations which are lethal at the larval-pupal transition are often related to mutations affecting essential mitotic functions (Gatti and Baker, 1989). The underlying reason is that the zygotic expression of many cell-cycle functions is not required for larval viability, because most cell-cycle activity ceases in the larvae (Shearn *et al.*, 1971; Gatti and Baker, 1989). With the onset of the cell-cycle in the imaginal discs the expression of these functions is required for viability. A systematic cytological screen of mapped mutants displaying a lethal phase at this transition plus degenerated imaginal discs, proved to be highly successful. In addition to those mutants known to be involved in the cell cycle (Ripoli *et al.*, 1985; Smith *et al.*, 1985; Gonzales *et al.*, 1988; Sunkel and Glover 1988) this screen revealed thirty mutants which were found to be employed in essential mitotic processes. Some mutants displayed severe defects in chromatin condensation and in others the cell cycle was arrested at certain phases.

These examples illustrate how fruitful the analysis of lethal genes can be. Two lethal genes, *A112* and *LB20*, are the focus of this thesis. These genes are located at the base of the X chromosome in division 19/20. Several lethal screens were initiated in this area with the intention of establishing the fine-structure of a small chromosome segment. A cytogenetic map of lethal and non-essential loci, established from the combined information on these various mutagenesis experiments, reveals this segment to be one of the most completely mutagenized intervals in the *D. melanogaster* genome (Lifschytz and Falk, 1968, 1969; Schalet and Lefevre 1973, 1976; Lifschytz and Yakobovitz 1978; Lefevre 1981; Kramers *et al.*, 1983; Eeken *et al.*, 1985; Zusman *et al.*, 1985; Lefevre and Watkins, 1986; Schalet 1986; Green *et al.*, 1987).

A112 (Lifschytz and Falk, 1968) and *LB20* (Schalet and Singer, 1971) were discovered in separate mutagenesis experiments which had been carried out to answer a number of questions:

(i) What point on Bridges' (1938) map of the salivary X chromosome corresponds to the euchromatic-heterochromatic junction as seen in the mitotic X chromosome?

(ii) Do ordinary sex-linked genes (those with no alleles on the Y chromosome) occur in the proximal heterochromatic region of the X chromosome?

(iii) With which salivary gland chromosome is the nucleolus associated?

(iv) Are bands in the most proximal region of the salivary gland X chromosome, section 20, equivalent in their genetic significance to those found in more distal regions, such as

section 19, particularly with regard to a one-to-one relationship of genes to bands?

The genetic mutagenesis experiments, conducted by Lifschytz and Falk (1968, 1969), concentrated on a specific segment of the X-chromosome. Wild-type males from the Qiryat Anarium stock were irradiated with a dose rate of 3200 Röntgen (R). Lethal alleles from the X chromosome were subsequently isolated using the balancer chromosome $sc^{si}\ B\ I_n\ S\ w^a\ sc^8$ (Base). The lethal genes were localised by mating females heterozygous for a lethal gene to two types of males. One type of male carried the Y*mal*+ chromosome which bears the proximal fragment of the X chromosome including the wild type alleles of *sw mal* and *su(f)*, (Schalet and Finnerty, 1968a; Chovnick *et al.*, 1969; Schalet and Lefevre, 1973). The second type of male carried the Y*w*+ chromosome, which bears the segment of the X with the normal alleles of *br, pn, w,* and *spl* (Lifschytz and Falk, 1968a and 1968b). Thus alleles bearing lethal mutations in these two regions could be detected. The aim of the experiment was to determine all of the functional units in this segment as a prerequisite to a later study concerned with the functional relationships between these units.

In total 413 chromosomes carrying X-ray induced lethal mutations were obtained by irradiating 5302 X-chromosomes. 42 of these mutant chromosomes were found to be viable when crossed to Y *mal*+. These mutant chromosomes, including the one containing *A112*, thus contained lethal mutations within the chromosome segment defined by Y *mal*+ (Lifschytz and Falk, 1968). By testing these for complementation, 20 units were defined (Lifschytz and Falk, 1968).

The first reference to *LB20* was made by Schalet and Singer (1971) when this locus was mentioned as one of 12 new mutant sites, which were added to the existing map of the X-chromosome. It was stated that this locus is a single functional unit (Schalet and Singer, 1971). This map also contained the information from the mutagenesis screen conducted by Lifschytz and Falk (1968). Using these data, a revised map was established, depicting 32 complementation groups in the segment defined by Y *mal*$^{+}$ (Schalet and Lefevre, 1976). The lethal genes *A112* and *LB20* were placed, as a pair with an ambiguous proximal-distal orientation, in the most proximal region of 19F (Schalet and Singer, 1971; Schalet and Lefevre, 1976). This position is close to the transition zone of euchromatin to heterochromatin which has always been difficult to map (Schalet and Lefevre, 1973; 1976; Lefevre, 1981). The final orientation, *A112* distal to *LB20*, was determined by Miklos *et al.* (1986). A cytogenetic map which illustrates the current view of the complementation groups in this region is presented in chapter 2 (Fig. 2.1).

The genes *A112* and *LB20* were also potentially interesting because they reside in a region that contains numerous putative neurogenic genes (Markov and Merriam 1977; Fischbach and Heisenberg, 1981; Kelly, 1983; Thomas and Wyman, 1984; Miklos *et al.*, 1987; Perrimon *et al.*, 1989). The lethality of some of these genes indicated that these have functions that are essential to survival. Although the region 19/20 was well characterised structurally, little was known about *A112* and *LB20* and nothing was known about the density of the transcription units in the region.

Only recessive lethal alleles have been found at the *A112* locus and the alleles that have been tested for complementation include: *A112* ^{A112}, *A112* ^{11P1}, *A112* $^{17\text{-}62}$, *A112* ^{GF314} and *A112* $^{8\text{-}1}$ (Lifschytz and Falk, 1968; Schalet and Lefevre, 1976; Miklos *et al.*, 1986; Perrimon, *et al.*, 1989; Lefevre, unpublished data). Mutations at this locus are lethal at the embryonic-larval interface. Alleles which have been analysed with germline clone analysis reveal no homozygous germline clones, indicating that mutations at this locus are germ cell lethal, ie. that the function of the *A112* gene is required for cell viability (Perrimon *et al.*, 1989).

As is the situation at the A112 locus, only lethal alleles have been found at the LB20 locus. Alleles at this locus tested for complementation include: *LB20* ^{LB20}, *LB20* ^{DA618} and *LB20* ^{C27} (Schalet and Singer, 1971; Schalet and Lefevre 1976; Perrimon *et al.*, 1989; Miklos *et al.*, 1986; Lefevre unpublished data). Each of these alleles is lethal during the larval-pupal stages. Germline clone analysis reveals that these mutations cause germ cell lethality (Perrimon *et al.*, 1989).

There is also some evidence that *LB20* might be a mitotic mutant. Mitotic chromosomes in brain squashes of material from two alleles have been analysed. The analyses of these alleles of *LB20* (*DA618*, by Maurizio Gatti, [personal communication] and *C27* by Masatoshi Yamamoto, [personal communication]) showed chromosome division arrested in metaphase with contracted metaphase chromosomes, a high mitotic index, frequent polyploid cells and no anaphase. Because mutations at the *LB20* locus are lethal at the larval-pupal transition, this gene might be essential in mitotic processes (Gatti *et al.*, 1989).

Some additional information is available regarding the general area in which the genes *A112* and *LB20* reside. There is a viable mutant, *short egg* (Wieschaus *et al.*, 1981), in which the shape of the egg is less elongated and more rounded. The mutation giving rise to this phenotype has been mapped within *Df(1)JA117* (Wieschaus and Miklos, unpublished data). Because the DNA deleted in this deficiency extends only from *A112* to *LB20* it is possible that the *short egg* mutation might represent a visible allele for one of these genes.

Finally, there is also evidence that a "collagen-like" gene might reside in the region. Initial *in situ* hybridisation experiments showed that a chicken collagen probe mapped to this region in 19F/20A on the X chromosome (Natzle *et al.*, 1982).

Much can be learned about these mutants isolated in previous genetic screens by applying molecular techniques. The complementation groups *A112* and *LB20* are particularly interesting in this respect for several reasons:

(i). both were isolated as lethal alleles in X-chromosomal mutagenesis experiments

(ii). apart from their lethality in early development, mutant alleles at both loci appear to have no visible phenotype, which makes them less amenable to classical genetic analysis

(iii). both genes are cell lethal in the ovary

(iv). they are both located in a region which is interesting because it contains numerous neurological genes and also because this region is at the transition zone between euchromatin and heterochromatin.

In division 19 the faintly staining euchromatin is present next to heavier staining β-heterochromatin of division 20 which appears as a diffuse mass compared to the distinct bands of α-heterochromatin. Little is known in detail about the transition zone of this region although, it contains approximately one and a half megabases of DNA. It has also been shown that the β-heterochromatin exclusively contains several classes of repetitive sequences, one of which is the Dr. D family of repetitive sequences (John and Miklos 1988). The gene density is very similar to that observed for the adjacent euchromatic region (Perrimon *et al.*, 1989).

This thesis investigates two lethal genes which cannot readily be analysed further with classical genetic tools. The main aims of the study were to define the structural boundaries of the *A112* and *LB20* complementation groups, assess the number of transcripts produced in the region and then to determine the DNA sequences of the two genes in order to gain some insight into the biological functions of the proteins they encode.

The molecular mapping of the X-chromosomal genetic complementation groups *A112* and *LB20* (using chromosomal rearrangements, contiguous loci and deletion mutants [Schalet & Lefevre 1976]) is described in chapter 2.

In the third chapter the region within the mapped genomic limits of *A112* and *LB20* is analysed in more detail by localising possible transcription units within the cloned genomic area. A preliminary transcription unit map of the *A112/LB20* genomic DNA region is established by comparing the size and intensity of

the hybridisation signals together with the developmental profiles of the transcripts detected using a series of contiguous labelled probes made from genomic DNA subclones of the region.

Another approach to map the transcriptional activity of an uncharacterized region of genomic DNA is to use probes from the region to screen cDNA libraries derived from RNA of appropriate tissues or developmental stages. Cross hybridization of cDNA sequences back to the genomic DNA then allows the localization of transcribed regions and the grouping of cDNA clones into putative transcription units. This approach is the basis for the sequencing analyses described in chapters 4 and 5.

Finally, *in vivo* functional analyses of genomic DNA to verify the location of the characterized genes by assaying the genomic DNA fragments for their ability to rescue the mutant phenotype is described in chapter 6.

Chapter 2

Localisation of the loci *A112* and *LB20*

2.1 Introduction

Genetic analyses in *D. melanogaster* have been highly successful in isolating gene networks, for example those involved in the cellular determination of the embryonic nervous system (Goodman *et al.*, 1984, Thomas *et al.*, 1984, 1988) and those controlling embryonic segmentation processes (Ingham *et al.*, 1985; Ingham 1988). The success of this approach in *D. melanogaster* is mainly due to the availability of lethal mutations at some loci, and the subsequent isolation and characterization of the wild type products of these loci.

The physical position of a gene locus can be determined in *D. melanogaster* either by *in situ* hybridization of gene probes to salivary gland polytene chromosome or by genetic crosses mapping deficiencies. Using these techniques, it is possible to map genes at the cytological level to within tens of kilobases or to within one or a few bands depending on the deficiencies that are available (Roberts, 1986). By contrast, the accuracy of mapping chromosomal aberrations in human mitotic chromosomes, for example, is still only to within tens of millions of bases (Roberts, 1986).

For molecular analyses, a *D. melanogaster* gene can be obtained using either a probe from another organism to screen a *D. melanogaster* library or, if the cytogenetic location of the gene is known, a chromosomal walk can be initiated. The walk can begin

via previously cloned sites in the vicinity of the gene or via microcloning, where a mini-library is cloned from a single band excised from a salivary gland chromosome (Pirrotta 1986).

Regardless of whether the approach used to clone a gene is via a cytogenetic location or via a known DNA sequence, the molecular map has to be aligned with the physical genetic map in order to identify all of the sequence corresponding to the gene. In practice a number of different approaches are usually necessary to establish the identity of a gene, including the use of deletion breakpoints, re-arrangements and insertional mutations.

This chapter describes the molecular mapping of the X-chromosomal genetic complementation groups *A112* and *LB20*. The complementation group *LB20*, discovered by Schalet and Singer (1971) and the complementation group *A112*, identified by Lifschytz and Falk (1968), were obtained as part of X-ray induced mutagenesis experiments carried out in order to establish the genetic fine structure of a part of the X-chromosome in *D. melanogaster*. Both *A112* and *LB20* were tested for complementation and mapped to salivary region 19F/20A on the X-chromosome (Schalet and Singer, 1971; Schalet and Lefevre, 1976) proximal to the complementation groups *small optic lobes (sol)* and *sluggish (slg)* and distal to *tumorous head (tuh-1)* and *extra organs (eo)* (Fig. 2.1). Their current order on the cytogenetic map was established by Miklos *et al.* (1986).

Knowledge of the cytogenetic location of the genes *A112* and *LB20* near the base of the X-chromosome allowed the selection of an entry point for a chromosomal walk in this region. *DCg2*, a *D.*

melanogaster clone which had been isolated using a chicken collagen probe, was shown to hybridize to DNA in the general area of the polytene chromosome containing the *A112* and *LB20* loci (Natzle *et al.*, 1982) and served as an entry point for a chromosomal walk (Miklos unpublished data). The restriction map of the part of this chromosomal walk that is relevant to this thesis is illustrated in figure 2.1. *DCg2* hybridizes to a 1.3kb *Hind* III fragment at position 0 to 1.3 in the chromosomal walk. From there the walk was extended in both directions for about 180kb.

E F A B C D E F A B C D
18 19 19 20
19
20
-15 -10 -5 0 5 10 15 20 25 30 35 40 45 50
E E E E E E E E E E E E E E E
H H H H H H H H H H H

Figure legend to Fig. 2.1

Fig. 2.1 The position of the chromosomal walk within the genome of *D. melanogaster* collated from the sources indicated.

The base of the X chromosome is presented at the top illustrating the sections 19 and 20 of the polytene bands (adapted from Lefevre, 1976). The black box below represents the relative position of the probe used as an entry point for the chromosome walk (George Miklos, personal communication). *In situ* hybridization of this probe to sections 19 and 20 of the polytene chromosome (Natzle *et al.*, 1982) is illustrated by an arrow (pointing up). Hybridization of this probe to a *Hind* III fragment included in the walk is indicated below with dashed lines (George Miklos, personal communication). The left hand side of the *Hind* III fragment is set as position O in the co-ordination system shown below the restriction enzyme map. From this *Hind* III fragment the walk was extended in both directions (George Miklos, personal communication). The restriction enzymes sites *Eco* RI and *Hind* III are indicated by (E) and (H) respectively.

Using genomic DNA subclones of this chromosomal walk, I have analysed mutants that are deficient for the loci of interest. The deletion mutants chosen were those that by genetic complementation tests separate the loci *A112* and *LB20* from each other or from the adjacent complementation groups: the deficiencies used were *Df(1)HM44* (Miklos *et al.*, 1986),

Df(1)GA104, *Df(1)JC77*, *Df(1)JA117*, *Df(1)S54*, *Df(1)JC4* (Lefevre 1981), *Df(1)16-129*, *Df(1)Q539*, *Df(1)17-257* and *Df(1)B57* (Schalet & Lefevre 1976) (Fig. 2.2.).

Figure 2.2

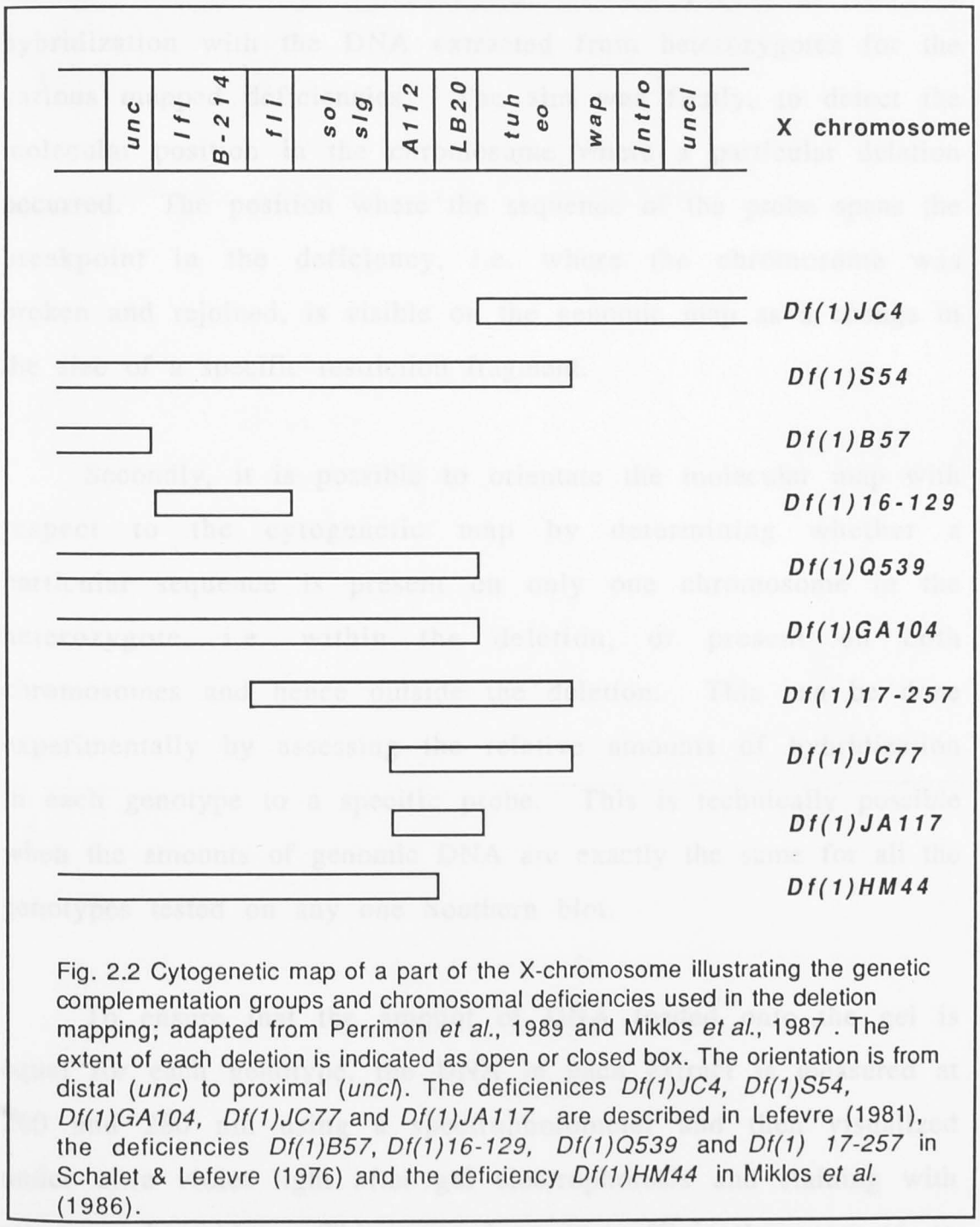


Fig. 2.2 Cytogenetic map of a part of the X-chromosome illustrating the genetic complementation groups and chromosomal deficiencies used in the deletion mapping; adapted from Perrimon *et al.*, 1989 and Miklos *et al.*, 1987 . The extent of each deletion is indicated as open or closed box. The orientation is from distal (*unc*) to proximal (*uncl*). The deficienices *Df(1)JC4, Df(1)S54, Df(1)GA104, Df(1)JC77* and *Df(1)JA117* are described in Lefevre (1981), the deficiencies *Df(1)B57, Df(1)16-129, Df(1)Q539* and *Df(1) 17-257* in Schalet & Lefevre (1976) and the deficiency *Df(1)HM44* in Miklos *et al.* (1986).

To align the physical map obtained from the complementation tests and deficiency analyses with the molecular map, the position of the deletion breakpoints were molecularly defined using genomic Southern blots (Fig. 2.3-2.12). This was done using the genomic DNA subclones derived from the chromosomal walk as radioactively labelled probes to test for hybridization with the DNA extracted from heterozygotes for the various mapped deficiencies. The aim was firstly, to detect the molecular position in the chromosome where a particular deletion occurred. The position where the sequence of the probe spans the breakpoint in the deficiency, i.e. where the chromosome was broken and rejoined, is visible on the genomic map as a change in the size of a specific restriction fragment.

Secondly, it is possible to orientate the molecular map with respect to the cytogenetic map by determining whether a particular sequence is present on only one chromosome in the heterozygote, i.e. within the deletion, or present on both chromosomes and hence outside the deletion. This can be done experimentally by assessing the relative amounts of hybridization in each genotype to a specific probe. This is technically possible when the amounts of genomic DNA are exactly the same for all the genotypes tested on any one Southern blot.

To ensure that the amount of DNA loaded onto the gel is equal for each genotype, the DNA in each extract is measured at 260 and 280 nm using a spectrophotometer and then visualized under ultra violet light after gel electrophoresis and staining with ethidium bromide. This procedure is sufficiently accurate to differentiate between a ratio of 2:1 and 1:1. The Southern blots

were then hybridized to a radioactively labelled probe of known origin and the relative intensities of the radioactive signals were compared for each genotype. If the relative amounts of hybridized probe are one to one then both chromosomes contain DNA which hybridizes to that probe, but if the ratio is two to one then only one chromosome in the heterozygote has the DNA homologous to that probe.

2.2 Materials and Methods

2.2.1 Bacterial strains and media

M13 bacteriophage and *pEMBL* plasmids (Dente *et al.*, 1983) were propagated in the *Escherischia coli* strain JM101. The bacteriophage λgt^{10} and λgt^{11} were propagated in *E. coli* strains JP777 or RY1090 respectively. The *E.coli* strain LE392 was used to propagate the bacteriophage λNM1149.

All bacterial strains were kept as glycerol-stocks (50% glycerol/ 50% LB-broth, LB-broth see below) at -70°C. Fresh bacterial colonies were obtained by taking a scrape from the frozen stock with a sterile loop and streaking the cells onto agar plates. The *E. coli* strain JM101 was generally streaked onto minimal agar plates unless it was used for transformation experiments (Hanahan 1983), when SOB plates were used. Media for minimal plates contained per litre final volume, 6g disodium hydrogen phosphate ($Na_2HPO_4 \cdot 7H2O$), 3g potassium dihydrogen phosphate (KH_2PO_4), 0.5g sodium chloride (NaCl) and 1g ammonium chloride (NH_4Cl) dissolved and adjusted to pH 7.4. 15g of bacto-agar was added and the solution was autoclaved and cooled to 60°C. The following sterilized solutions were added: 10ml of 20% glucose, 1ml of 0.1M calcium chloride ($CaCl_2$), 1ml of 1M magnesium sulphate ($MgSO_4$) and 50µl of thiamine at a concentration of 100µg/ml. The media for SOB plates contained 20g bacto-tryptone, 5g bacto-yeast and 0.5g NaCl dissolved in 950ml of distilled water to which was added 10ml of 250mM potassium chloride (KCl). The pH was adjusted to

pH 7 and 15g of Bacto-agar was added, the solution was autoclaved. Just before pouring the plates 5ml of sterile 2M magnesium chloride ($MgCl_2$) was added. Other *E. coli* strains were streaked onto LB plates (see below).

Liquid cultures of *E. coli* were grown in Luria-Bertani medium (LB medium) which contained per litre: 10g Bacto-tryptone, 5g yeast extract and 10g NaCl. LB-plates were prepared by adding 15g of bacteriological agar (Difco) to the above medium and autoclaving.

Selection for *E.coli* JM101 cells successfully transformed with *pEMBL* constructs was performed on Ampicillin plates. To prepare Ampicillin plates the LB solution was autoclaved and cooled to 50°C before Ampicillin was added to a final concentration of 100µg per ml.

M13 bacteriophage were plated onto LB-plates in an overlay of LB-top-agar (0.7% agarose in LB-medium). In order to allow the blue-white colour selection which marked an insertion into the *lac* Z gene, 50µl of 2% X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) and 2.5mg of IPTG (Isopropylthio-β-galactoside) was added per plate.

2.2.2 *Drosophila* stocks and media

All *Drosophila* strains used were obtained from Dr. G.L.G. Miklos (Research School of Biological Sciences, Australian National University, Canberra). X chromosomal deficiencies were balanced with one of the following balancer chromosomes FM6, FM7 or Binsn. Flies were reared at 20°C on standard maize meal,

molasses, yeast medium. 10g Bacto Agar, 15g sucrose, 40g yeast, 40g malt, 30ml Karo corn syrup and 10g soya flour were boiled in 1 l water for 20 minutes. The medium was cooled to 50°C before 4.5ml of propionic acid and 9ml of a 10% Nipagen solution in ethanol was added.

2.2.3 DNA preparation

(i) Small scale preparation of plasmid DNA

Mini-preparations of plasmid DNA were based on the alkaline lysis method (Birnboim and Doly 1979, Sambrook *et al.* 1989). A single bacterial colony was transferred into 2ml of LB-medium, containing the appropriate antibiotics and incubated in a shaker overnight at 37°C. The cells were harvested in an Eppendorf tube by centrifugation, aspirated and resuspended in 100µl solution 1 (solution 1: 50mM glucose, 25mM tris-HCl pH 8.0, 10mM EDTA pH8.0) containing 5mg/ml lysozyme. After 5 minutes incubation at room temperature 200µl solution 2 was added (solution 2: 0.2N NaOH, 1% sodium dodecyl sulphate (SDS) and carefully mixed by inverting the tube. The tube was stored for 10 minutes on ice before 100µl solution 3 was added (solution 3: 60 ml of 5M potassium acetate, 11.5ml of glacial acetic acid, 28.5ml of double distilled water). The viscous lysate was stored on ice for at least 10 minutes and then centrifuged. The supernatant was phenol extracted and the DNA precipitated with 2 volumes of ethanol. The DNA was sedimented by centrifugation, the pellet air-dried and dissolved in 50µl TE (TE: 10mM tris-HCL pH8, 1mM EDTA). To test the DNA preparation usually 1µl of DNA was analysed with restriction endonucleases.

(ii) Large scale preparation of plasmid DNA

Large quantities of plasmid DNA were prepared essentially as described above with the following modifications. 100-200ml LB-medium, containing the appropriate antibiotics, were inoculated with a single bacterial colony to grow overnight. The cells were harvested by centrifugation at 7,000 rpm, drained well and resuspended in 10ml solution 1. The lysis was completed by adding 20ml of solution 2 and 10ml of solution 3 with a 10 minute incubation on ice after each step. The precipitate of cell debris was removed by centrifugation at 8,000 rpm. The nucleic acids were precipitated with 0.6 volumes of propan-2-ol and pelleted by centrifugation at 8,000rpm. The pellet was drained well and resuspended in 2.8ml TE. To this solution was added 2.8g cesium chloride (CsCl) and 280µl ethidium bromide. In order to band the DNA on a cesium chloride gradient, the mixture was transferred to Beckman Quick-seal polyallomer tubes and centrifuged at 100,000 rpm for 4-16 hours at 25°C. The centrifugation was then extended for 1 hour at reduced speed (72,000 rpm) to relax the gradient. The plasmid DNA was located in the lower of two bands and was collected with a syringe. The ethidium bromide was removed from the plasmid solution with 6 or more extractions using equal volumes of CsCl saturated propan-2-ol. The CsCl was removed by dialysis overnight against TE with several changes of TE buffer.

(iii) Preparation of bacteriophage lambda DNA

200µl bacterial cells were taken from an overnight liquid culture, 200µl of 0.1M $CaCl_2$, 0.1M $MgCl_2$ were added and this culture was inoculated with bacteriophage λ. The correct amount of bacteriophage was determined empirically; usually 0.5µl, 1µl and 10µl of a particular bacteriophage suspension was used. The

cultures were incubated for 30 minutes at 37°C and then used to inoculate 100ml of medium each and incubated overnight at 37°C with vigorous shaking. The following morning 1ml of chloroform was added to each flask and the cultures were shaken another 15 minutes. Of the three flasks, the two with better lysis were chosen for the preparation. The 200ml lysates were centrifuged at 5,000 rpm for 20 minutes to remove the bacterial debris. Centrifugation of the supernatant for 4 hours at 11,000 rpm pelleted the bacteriophage. The sedimented bacteriophage were resuspended in 4.3ml of SM buffer (SM: per 1 litre: 5.8g NaCl, 2g $MgSO_4 \cdot 7H_2O$, 50ml 1M tris-HCl pH7.5, 5ml 2% gelatin). 3.4g CsCl was dissolved in 4ml of this suspension, the solution was transferred into a Beckman ultraclear centrifuge tube. The bacteriophage were banded by centrifugation at 36,000 rpm for 16 hours in a SW 55Ti rotor and collected by side puncture. The CsCl was removed by dialysis overnight with several changes of TE buffer.

The bacteriophage, containing the DNA, were incubated with Na_2EDTA at a final concentration of 0.1M for 10 minutes at 60°C. After adding Pronase to a final concentration of 1mg/ml and SDS to a final concentration of 0.5% the solution was incubated for a further hour at 37°C. The DNA was phenol-chloroform extracted and dialysed overnight with several changes of TE-buffer.

(iv) Preparation of genomic DNA

Genomic DNAs isolated from the various heterozygotes used in the mapping analysis were supplied by Dr. G.L.G. Miklos. The DNA was extracted using the cesium chloride method (Sambrook *et al.*, 1989).

2.2.4. Digestion of DNA with restriction enzymes.

Restriction endonucleases were obtained from Amersham, Boehringer Mannheim and New England Biolabs. One unit of restriction enzyme was used to digest approximately 2μg DNA in a Tris buffered solution at pH7.5 containing bovine serum albumin (0.2mg/ml final concentration), dithiothreitol (DTT, 1mM final concentration), $MgCl_2$ (10mM final concentration) and NaCl (usually 100mM). The concentration of NaCl and Tris-buffer were chosen as recommended by the supplier. Digestions were generally carried out for a period of 2 hours at 37°C, unless recommended differently by the supplier. DNA length standards were either generated by digestion of Lambda DNA with the restriction endonuclease *Hind* III or obtained commercially in the form of a "1Kb ladder" (from Bethesda Research Laboratories).

2.2.5 Purification of DNA

(i) Purification of DNA with phenol-chloroform

Nucleic acids were usually purified by phenol-chloroform extraction. Re-distilled phenol was saturated with distilled water. Working aliquots were buffered with 100mM Tris (pH8) by adding equal amounts of 100mM Tris (pH8), mixing and aspirating the aqueous phase. This was repeated until the phenolic phase reached pH 7.8. The extraction was performed by adding half a volume of phenol (pH7.8) and half a volume of chloroform to the DNA solution, mixing thoroughly and separating the two phases by centrifugation for 3 minutes in an Eppendorf centrifuge. The

upper phase was then usually precipitated with ethanol (see Section 2.2.6).

(ii) Purification of DNA by electroelution

To purify DNA fragments, sufficient DNA was digested with the appropriate restriction enzyme and electrophoresed on an agarose gel to yield 2-5μg DNA of the required fragment. The DNA was visualized under ultraviolet light and the fragment in question was excised and placed in a dialysis bag. The dialysis bag was boiled prior to use in TE for 1 minute and rinsed in double distilled water. TE contained 10mM Tris pH8 and 1mM ethylene diamino tetra acidic acid (EDTA). The dialysis bag, contained the DNA fragment in a solution of 0.5 x tris-borate buffer (TBE). 10 x TBE buffer pH8.3 contained per litre 108g tris-base, 55g borate, 9.3g $Na_2EDTA \cdot H_2O$. The dialysis bag was placed in an electrophoresis tank and submerged in 0.5 x TBE perpendicular to the current flow. Electrophoresis was carried out at 50mA for 2-4 hours. The DNA was detached from the wall of the bag by reversal of the current for 1 minute. The DNA was transferred to an Eppendorf tube and precipitated with ethanol (see Section 2.2.6.).

2.2.6 Ethanol precipitation

Ethanol precipitation of nucleic acids was carried out by adding 1/10 volume of 3M sodium acetate pH5.5 to the DNA solution, mixing briefly and adding 2.5 volumes of 96% ethanol. Small quantities (less than 1μg) were precipitated at -20°C overnight; amounts greater than 1μg were kept at room temperature for 15 minutes. The DNA was sedimented by

centrifugation at 12,000 rpm for 15 minutes, aspirated and washed with 70% ethanol, drained and dried.

2.2.7 Ligation of DNA

DNA-fragments with cohesive ends were ligated in 50mM Tris-buffer pH7.6, 10mM $MgCl_2$, 1mM DTT, 5mM Adenosine-5'-triphosphate (ATP) and 1mM Spermidine with 0.1 Weiss unit of T_4 ligase. Typically vector and target DNA were ligated together in a molar ratio of 1:3 with a total DNA concentration of 0.2µg in a total volume of 10 µl at 4°C overnight.

2.2.8 Preparation and transformation of competent cells

For most experimental procedures competent cells were prepared based on the calcium chloride-method, according to the protocol in Maniatis *et al.* (1982). For experimental procedures where the transformation efficiency had to be higher than 10^8 transformed colonies per microgram of DNA, cells were made competent following the method developed by Hanahan and Meselson (1983), as described in chapter 4.2.

Successful recombinants in the transformation experiments were normally identified using a colour reaction. All plasmid vectors used contained the regulatory sequences and the coding information for the first part of the β-galactosidase gene (*lacZ*). The product of the *lacZ* gene was complemented with the defective β-galactosidase of the host cell to form an active protein. The activation of β-galactosidase was the basis for the formation of blue plaques or colonies. Colourless plaques indicated that the β-

galactosidase gene had been inactivated due to the insertion of foreign DNA (Sambrook *et al.*, 1989).

2.2.9 Analysis of DNA

Localization of specific sequences within genomic DNA was normally accomplished by the transfer techniques described by Southern (1975). Between 0.2µg and 1µg DNA, in the case of plasmid DNA and 5 to 10µg, in the case of genomic DNA, was digested with the appropriate restriction endonucleases and separated on an agarose gel.

(i) Agarose gel electrophoresis

Depending on the experimental procedure, agarose gels between 30ml and 300ml were cast in agarose gel beds (Bethesda Research Laboratories). The agarose was melted in 1xTBE buffer containing 1µg/µl ethidium bromide. Separation of low molecular weights was undertaken with 2% agarose, higher molecular weights were separated in 0.8% agarose. The digested DNA samples were mixed with one tenth volume of loading dye (25% Ficoll type 400, 0.25% bromophenol blue) and loaded into the gel slots. The electrophoresis was carried out in a horizontal gel apparatus (Bethesda Research Laboratories) in 1xTBE running buffer at various electrical current densities. The DNA was visualized under ultra violet light and photographed with Polaroid film (type 55). After visualization the DNA was transferred onto a membrane.

(ii) Transfer of DNA from agarose gels onto membranes

Prior to the transfer the DNA was denatured by shaking the gel for 30 minutes in 0.8M NaCl, 0.4M NaOH, neutralized for 40

minutes in 1.5M NaCl, 0.5M Tris pH8. The DNA was then transferred onto Hybond-N (Amersham) using the capillary transfer method as described in Maniatis (1989). After the transfer was completed, the Hybond-N filter was washed in 2xSSC, air-dried and exposed to UV light for 5 minutes to fix the DNA to the filter. The filter was wrapped in plastic film (Gladwrap) and stored at room temperature until it was used for hybridization to radioactively labelled probes.

(iii) Hybridization of DNA to ^{32}P-labelled probes

The filters were prehybridized at 65°C in plastic bags containing approximately 0.2ml hybridization solution per square centimetre of Hybond-N membrane (hybridization solution: 5xSSC, 0.2%BSA, 0.2%Ficoll, 0.2%polyvinyl pyrolidone, 0.1%SDS, 50µg/ml single stranded salmon sperm DNA). After 1 to 4 hours the radioactive probe was denatured by heating for 5 minutes at 100°C and added to the hybridization solution. The plastic bag was sealed carefully and placed in a shaking water bath. The DNA was allowed to hybridize to the probe for a minimum time of 16 hours at 65°C. The filter was washed 3 times for half an hour in 2xSSC, 0.1% SDS at 65°C, wrapped in Gladwrap and exposed to a X-ray film (Kodak X-Omat XAR-5) together with an intensifying screen (Dupont Lighting Plus) overnight at -80°C.

(iv) Radioactive labelling of DNA fragments

DNA was labelled with α-^{32}PdCTP (Amersham) following the method of Feinberg and Vogelstein (1983). 50ng-150ng of target DNA in a volume of 35µl was denatured at 100°C for 10 minutes. 10µl of oligolabelling buffer was added. (oligo-labelling buffer: OLB was prepared from Solution A, B and C at the

ratios 100:250:150. Solution A contained 1ml 1.25M tris-HCl pH8.0, 0.125M $MgCl_2$, 18µl β-mercaptoethanol, 5µl of each dNTP at a concentration of 100mM. Solution B was 2M Hepes buffer pH6.6. Solution C was Pharmacia hexanucleotide primers at a concentration of 90 O.D. units/ml.) To the buffered DNA was also added 2µl BSA (at a concentration of 10 mg/ml), 2-5µCi of α-^{32}PdCTP, 2 units of the large fragment of *E.coli* DNA polymerase I (Amersham) and distilled water to 50µl. The oligolabelling reaction was usually incubated at room temperature overnight. To remove unincorporated nucleotides, the labelled DNA was precipitated by adding 1.2µl of 0.25M spermine, incubated on ice for 15 minutes and pelleted for 10 minutes in a centrifuge. The incorporation rate was estimated by measuring the radioactivity of the pellet and the supernatant, assuming that the pellet contained mostly incorporated nucleotides. On average the incorporation was 80%. The pellet was resuspended in 200µl of 10mM Na_2EDTA, 0.5% SDS and used immediately for hybridization.

2.3 Results

2.3.1 Mapping of *Df(1)HM44*

Df(1)HM44 is a large deficiency which has been shown to have its proximal breakpoint between *A112* and *LB20*, thus separating the two loci (Miklos *et al.*, 1986). The analysis of flies heterozygous for *Df(1)HM44*, to test whether the sequences of a probe used are present on both chromosomes or only on the wild-type chromosome, should therefore provide important clues on the alignment of the molecular map with the cytogenetic map. (Throughout this thesis the orientation of the cytogenetic and molecular maps are drawn placing proximal to the right and distal to the left.)

In figure 2.3 *Df(1)HM44* is compared to the deficiencies *Df(1)JC4*, *Df(1)GA104* and *Df(1)JA117*. Cytogenetically, *Df(1)HM44* and *Df(1)GA104* both uncover the *A112* locus, i.e. alleles at the *A112* locus are lethal in hemizygotes with *Df(1)HM44* and *Df(1)GA104*. The *LB20* locus is uncovered only by *Df(1)GA104*, and not by *Df(1)HM44* (Lefevre, 1981; Kramers *et al.*, 1983; see also Fig. 2.2), i.e. alleles at the *LB20* locus are lethal in hemizygotes with *Df(1)GA104*, and viable with *Df(1)HM44*. *Df(1)JC4*, the third deficiency used, has its distal breakpoint between the loci *LB20* and *tumerous head / extra organs* (*tuh-1/eo*), and hence does not impinge on *LB20* (Lefevre, 1981; Fig. 2.2).

Figure 2.3

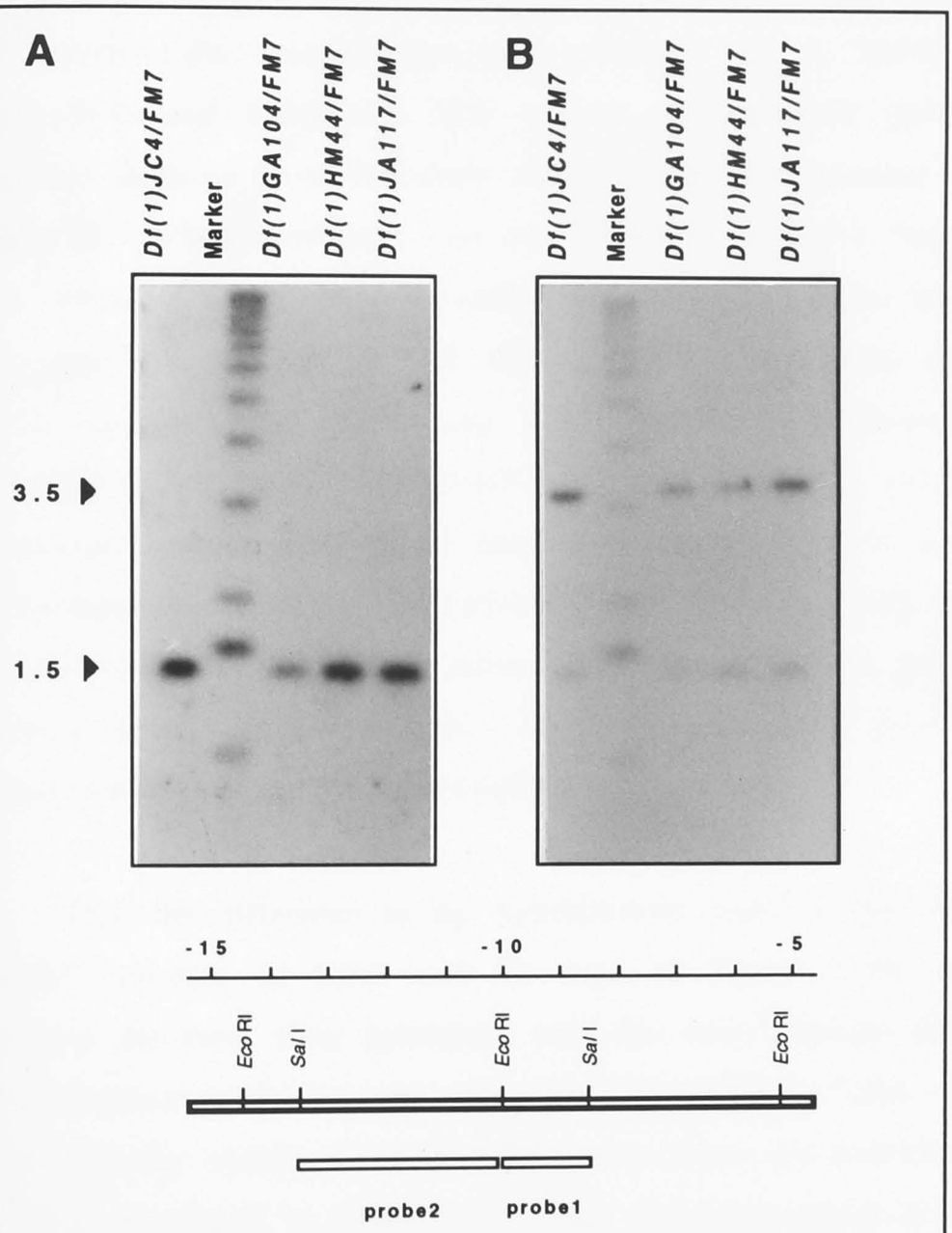


Fig. 2.3 Southern analysis of *Df(1)HM44*
Genomic DNA of the genotypes *Df(1)JC4/FM7*, *Df(1)GA104/FM7*, *Df(1)HM44 /FM7* and *Df(1)JA117/FM7* was cleaved with the restriction endonucleases *Eco* RI and *Sal* I, transferred to a membrane and hybridized with probe 1, the -8.5 to -10 *Eco* RI/*Sal* I fragment (panel A) and probe 2, the -10 to -13.5 *Eco* RI/*Sal* I fragment (panel B). Note (in panel A) the equivalent intensities of the radioactive signals at 1.5kb for the genotypes *Df(1)JC4/FM7*, *Df(1)HM44 /FM7* and *Df(1)JA117/FM7* and (in panel B) the lower intensity of the signal corresponding to *Df(1)HM44 /FM7* at 3.5kb in comparison to the signals of *Df(1)JC4/FM7* and *Df(1)JA117/FM7* at 3.5kb.

In order to pin-point the molecular breakpoint of *Df(1)HM44*, DNA from the heterozygotes *Df(1)HM44/FM7*, *Df(1)GA104/FM7* and *Df(1)JC4/FM7* was digested in equal amounts with restriction endonucleases, transferred to a membrane (Hybond-N) and hybridized with various ^{32}P-labelled genomic fragments isolated from subclones derived from the genomic walk (Fig. 2.3). When hybridized with the 1.5kb *Eco* RI/*Sal* I fragment from -10 to -8.5, one band is visible at 1.5kb (Fig. 2.3A) in each genotype. The intensities of the hybridization signals appear similar in the lanes containing DNA from the heterozygotes *Df(1)HM44/FM7* and *Df(1)JC4/FM7* indicating that these genotypes contain an approximately equal number of copies. In comparison the hybridization signal in *Df(1)GA104/FM7*, corresponding to the band at 1.5kb, is clearly less intense, indicating that this genotype contains fewer copies of this fragment than the genotypes *Df(1)JC4/FM7* and the *Df(1)HM44/FM7*.

That the difference in the hybridization signal is not due to unequal amounts of DNA can be seen in figure 2.3B, which illustrates the same filter hybridized with the distal adjacent *Eco* RI /*Sal* I fragment (-13.5 to -10). One band is visible at 3.5kb, and a second, faintly visible, band at 1.5kb stems from the hybridization described previously in figure 2.3A. The signal intensities at 3.5kb are equivalent in the lanes containing DNA from *Df(1)HM44/FM7* and *Df(1)GA104/FM7* heterozygotes, but these signals are less intense than the signal in the lane containing DNA from the *Df(1)JC4/FM7* heterozygote. Based on the assumption that the intensity of the hybridization signal is directly related to the number of copies of homologous DNA, and therefore to the number of wild-type X-chromosomes, it can be concluded that the

sequences homologous to the -13.5 to -10 *Eco* RI /*Sal* I fragment (probe 2, Fig. 2.3B) are present in one copy equivalent in the *Df(1)HM44/FM7* and *Df(1)GA104/FM7* heterozygotes. This one copy equivalent represents DNA in the wild-type chromosome which is homologous to the probe. The two copies equivalent present in the *Df(1)JC4/FM7* heterozygote correspond to the sequences present in the wild-type chromosome and in the chromosome bearing the deficiency.

A similar argument applies to the sequences homologous to the -10 to -8.5 fragment (probe 1, Fig. 2.3A). They are present in only one chromosome, the wild-type chromosome, in the *Df(1)GA104/FM7* heterozygote but they are present in two chromosomes in the *Df(1)HM44/FM7* and *Df(1)JC4/FM7* heterozygotes. Thus the proximal breakpoint of *Df(1)HM44* has to be proximal to the -13.5 to -10 *Eco* RI/*Sal* I fragment (probe 2, Fig. 2.3B), yet it has also to be distal to the -10 to -8.5 *Eco* RI/*Sal* I fragment (probe 1, Fig. 2.3A), which places it close to the *Eco* RI restriction site at -10.

There is no evidence in figure 2.3 that any of the genotypes give a fragment of altered size when probed with either the -13.5 to -10 *Eco* RI/*Sal* I fragment or the -10 to -8.5 *Eco* RI /*Sal* I fragment. It may be that the new *Eco* RI /*Sal* I fragment created in the deficiency bearing chromosome is of exactly the same size as the *Eco* RI /*Sal* I fragment from the wild-type chromosome and hence not distinguishable. A second possible explanation is that the breakpoint in *Df(1)HM44* is very close to the *Eco* RI restriction site at -10, and hence the radioactively labelled sequences of the

Figure 2.4

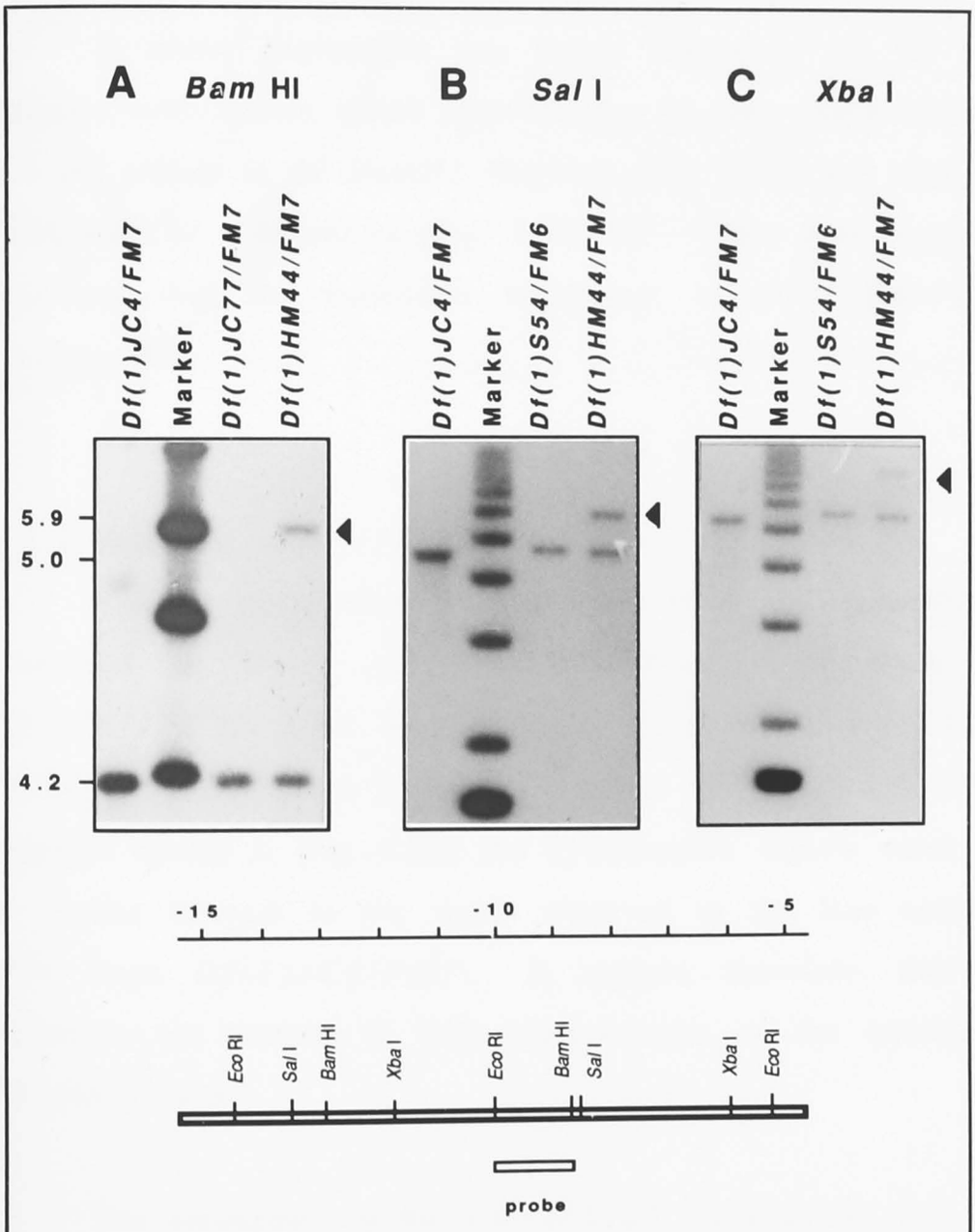


Fig. 2.4 Southern analysis of the breakpoint in *Df(1)HM44*.
Southern transfers of genomic DNA of various genotypes (as indicated) was cut with either *Bam* HI (panel A), *Sal* I (panel B) or *Xba* I (panel C), and hybridized to the 1.4 kb *Eco* RI/*Bam* HI fragment from position -8.6 to -10. The restriction map indicating the position of the genomic fragment used as a probe is illustrated below, the co-ordinate system is according to Fig. 2.1. Each panel containing *Df(1)HM44* DNA reveals a band additional to the expected signals present in the control DNAs, confirming that the probe spans the breakpoint.

probe that are homologous to the altered fragment are too small to be detected.

In either explanation one would expect to see an altered fragment with probes which span the *Eco* RI site. Such changes in size are present in the *BamH1* fragment (Fig. 2.4A) and also in the *Sal* I and *Xba* I fragments (Fig. 2.4B, C). These data support the argument that the molecular breakpoint of *Df(1)HM44* is at position -10.

2.3.2 Mapping of *Df(1)JA117*

Another observation can be made from the autoradiograph presented in figure 2.3. In Southern blots made from *Df(1)JA117/FM7* DNA hybridized with either the -13.5 to -10 *Eco* RI/*Sal* I fragment (probe 2, Fig. 2.3B) or the -10 to -8.5 *Eco* RI/*Sal* I fragment (probe 1, Fig. 2.3A) the hybridization signals seem to be of similar strength to the signal observed in the lane containing DNA from *Df(1)JC4/FM7*. It appears, therefore, that these fragments are present in both chromosomes of the heterozygote *Df(1)JA117/FM7*.

The breakpoint in the *Df(1)JA117* chromosome has to be either distal or proximal to the -13.5 to -8.5 region. Since heterozygotes *LB20/Df(1)JA117* and *A112/Df(1)JA117* are lethal, but only *A112/Df(1)HM44* is lethal, it seems most likely that the breakpoints of *Df(1)JA117* are proximal to the -13.5 to -8.5 region.

Figure 2.5 illustrates the hybridization of the proximal adjacent fragments to DNA extracted from flies heterozygous for

Figure 2.5

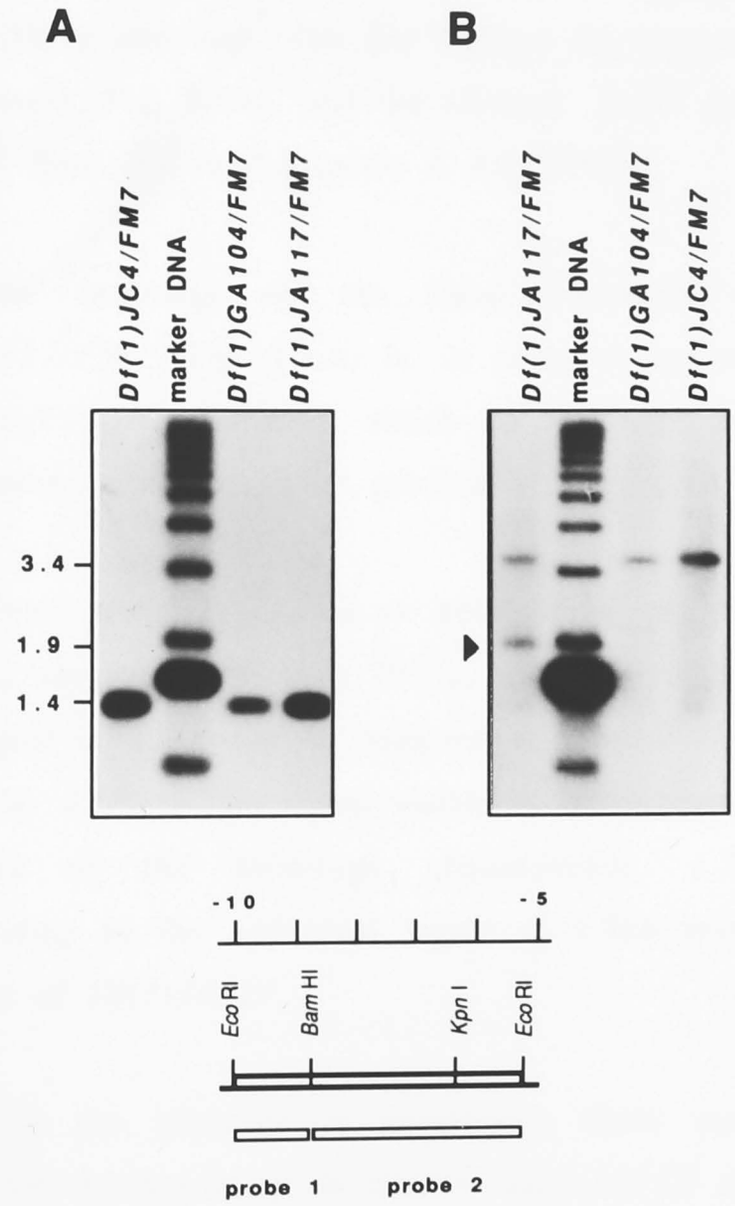


Fig. 2.5 Mapping of the distal breakpoint in *Df(1)JA117*.
Southern transfers of genomic DNA cleaved with *Eco* RI and *Bam* HI were hybridized to two radioactively labelled probes. Panel A: Autoradiogram of the hybridization to probe 1, the 1.4kb *Eco* RI/*Bam* HI fragment from -10 to -8.6. Panel B: Autoradiogram of the hybridization with probe 2, the 3.4kb *Eco* RI/*Bam* HI fragment from -8.6 to -5.2. The co-ordinates correspond to those in Fig. 2.1. Molecular sizes are indicated in kb. The novel fragment at 1.9kb indicates the breakpoint of *Df(1)JA117*.

Df(1)JA117/FM7 in comparison to DNA made from flies heterozygous for *Df(1)JC4/FM7* and for *Df(1)GA104/FM7*. The DNA samples were double digested with *Eco* RI and *Bam* HI, transferred to a membrane and hybridized with two ^{32}P radioactively labelled probes. These were the 1.4kb *Eco* RI/*Bam* HI fragment from -10 to -8.6 (probe 1, Fig. 2.5A) and the adjacent 3.4kb *Eco* RI/*Bam* HI fragment from -8.6 to -5.2 (probe 2, Fig. 2.5B).

The intensity of the band observed at 1.4kb in *Df(1)JA117/FM7* (Fig. 2.5.A) is, as expected, approximately twice that in *Df(1)GA104/FM7* which is in agreement with the hybridization pattern observed previously (Fig. 2.3).

When probe 2 is used to hybridize to an equivalent filter containing samples of the same DNAs, also cut with *Eco* RI and *Bam* HI, a signal of 1.9kb can be observed in addition to the signal at 3.4kb (Fig. 2.5B). The signal visible at 3.4kb corresponds to the sequences of the wild-type chromosome. The fragment corresponding to the additional signal at 1.9kb reveals the distal breakpoint of *Df(1)JA117*.

With the intention of confirming these results and pin-pointing the breakpoint of deficiency *Df(1)JA117* more precisely, another set of Southern blots was made using the same DNA samples but cut with different restriction endonucleases (Fig. 2.6). Two probes were isolated, the 3.6kb *Eco* RI /*Kpn* I fragment from position -10 to -6.4, which was expected to exhibit the breakpoint, and the 1.2kb *Eco* RI /*Kpn* I fragment from position -6.4 to -5.2. Figure 2.6A shows the Southern blot of genomic DNA double digested with *Eco* RI and *Kpn* I and hybridized with the 3.6kb *Eco*

Figure 2.6

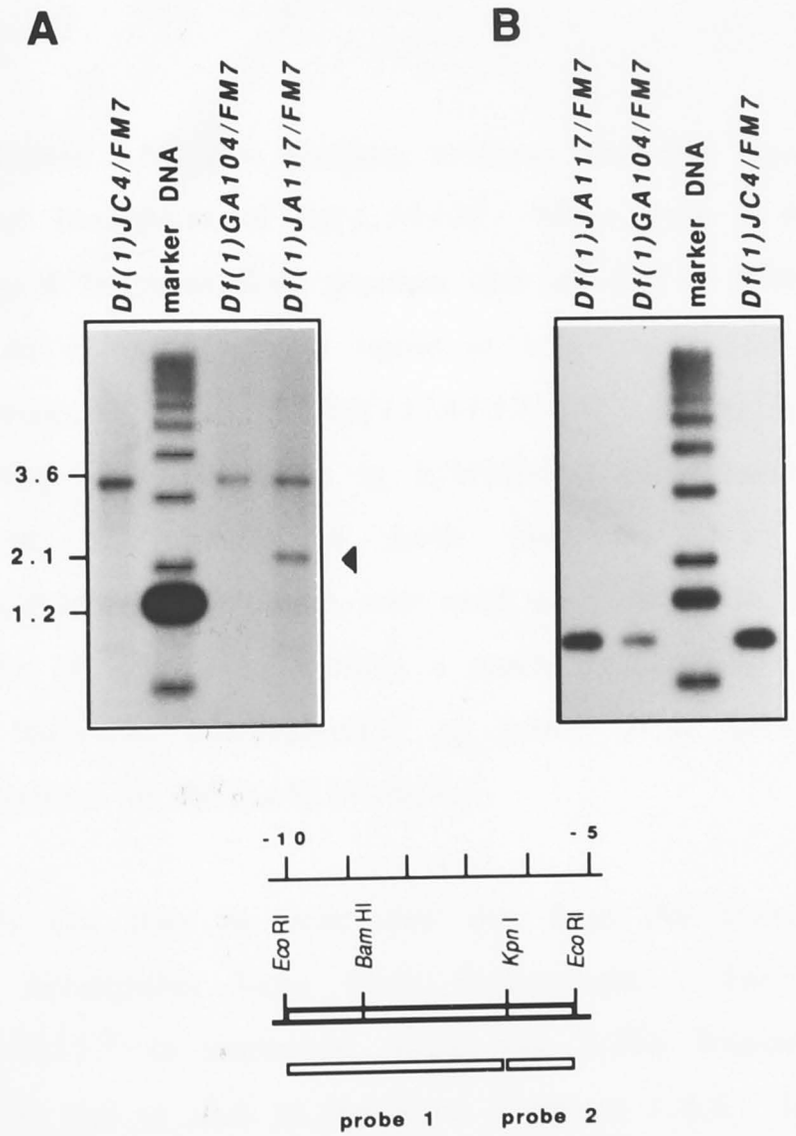


Fig. 2.6 Mapping of the proximal breakpoint of *Df(1)JA117*.
Southern transfers of genomic DNA cleaved with *Eco* RI and *Kpn* I were hybridized to two radioactively labelled probes. Panel A: Autoradiogram of the hybridization to probe 1, the 3.6kb *Eco* RI/ *Kpn* I fragment from -10 to -6.4. Panel B: Autoradiogram of the hybridization with probe 2, the 1.2kb *Eco* RI/ *Kpn* I fragment from -6.4 to -5.2. The co-ordinates correspond to those in Fig 2.1. Molecular sizes are indicated in kb. A novel fragment at 2.1kb is visible, indicating the breakpoint of *Df(1)JA117*. Also, the proximal probe 2 (panel B) is in two copies in *Df(1)JA117* as the signal intensities show, suggesting this breakpoint is the proximal breakpoint of *Df(1)JA117*.

RI/*Kpn* I fragment from position -10 to -6.4 (probe 1). Two bands are visible, one at 3.6kb which is present in all 3 lanes, and one at 2.1kb which is only present in the lane containing DNA from *Df(1)JA117/FM7*. This band represents the breakpoint of *Df(1)JA117*.

Figure 2.6B also provides evidence that this breakpoint is the proximal breakpoint of *Df(1)JA117*. When probe 2, the 1.2kb *Eco* RI/*Kpn* I fragment from position -6.4 to -5.2, is hybridized to the filter the intensity of the signal at 1.2kb is similar in the lanes containing DNA from *Df(1)JA117/FM7* and *Df(1)JC4/FM7*, indicating that the probe is hybridizing to approximately equal numbers of copies in both samples. In comparison *Df(1)GA104/FM7*, which was used as a reference for the signal intensity of one copy, exhibits a much weaker signal. Hence the DNA fragment corresponding to probe 2 is present on both chromosomes in *Df(1)JA117/FM7*.

It can thus be concluded that both the proximal and the distal breakpoint have been determined. The deletion in *Df(1)JA117* is contained within the 2.2kb fragment from the *Bam* HI site at -8.6 to the *Kpn* I site at - 6.4. It can also be concluded that the size of the deletion is approximately 1.5kb, since both new fragments are approximately 1.5kb smaller than the fragments corresponding to the wild-type chromosome.

This was confirmed by digesting genomic DNA of *Df(1)JA117/FM7* separately with *Bgl* II, *Sal* I and *Bam* HI and hybridizing the Southern transfer of each digest with the -10 to -5.2 *Eco* RI fragment which encompasses the deletion. Each of

Note:
The molecular weight markers used on this gel are not included in the photograph. The relevant sizes are indicated. The restriction map for the endonuclease *SalI* is included in figure 2.4.

the restriction enzymes used generated a fragment approximately 1.5kb smaller than the wild-type band (Fig. 2.7).

Figure 2.7

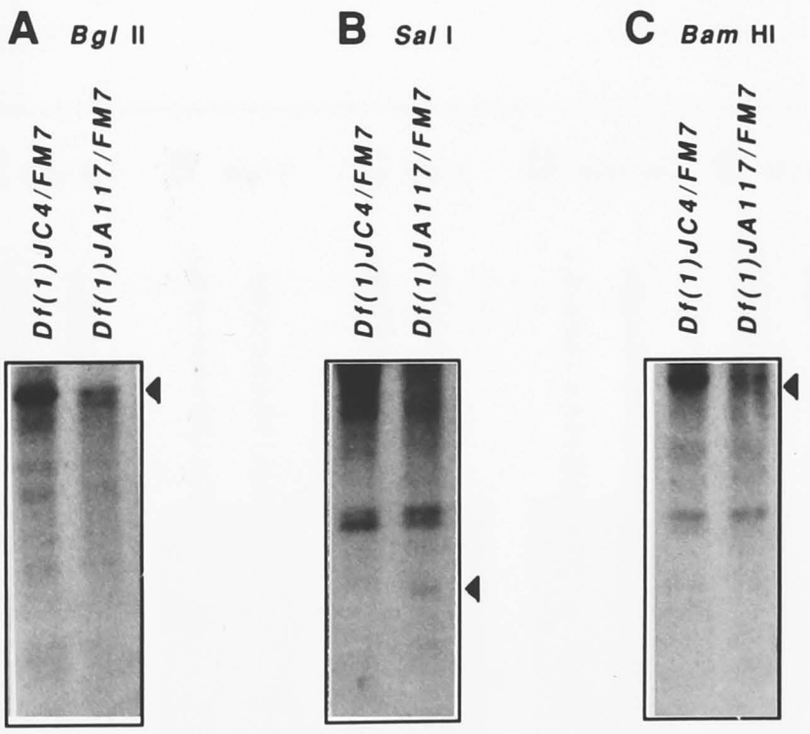


Fig. 2.7 Localization of the deletion in *Df(1)JA117*.
Genomic DNAs from the genotypes *Df(1)JA117/FM7* and *Df(1)JC4/FM7* were cleaved with *Bgl* II (A), *Sal* I (B) and *Bam* HI(C), separated by gel electrophoresis, blotted and hybridized with a labelled *Eco* RI fragment (-10 to -5.2 according to the co-ordinate system of Fig. 2.1). All novel fragments (10kb in A, 3kb inB and 9kb in C) are found to be approximately 1.5kb smaller than the wild-type band indicating that the mutation in *Df(1)JA117* is a deletion of 1.5kb.

2.3.3 Mapping of *Df(1)Q539*

The chromosome *Df(1)Q539* carries a deletion that uncovers all of the complementation groups from *varied outspread* (*vao*) to *LB20* (Schalet and Lefevre 1976). The proximal breakpoint of *Df(1)Q539* was analysed in a similar manner to that described previously for *Df(1)HM44* and *Df(1)JA117* (Fig. 2.8).

Figure 2.8

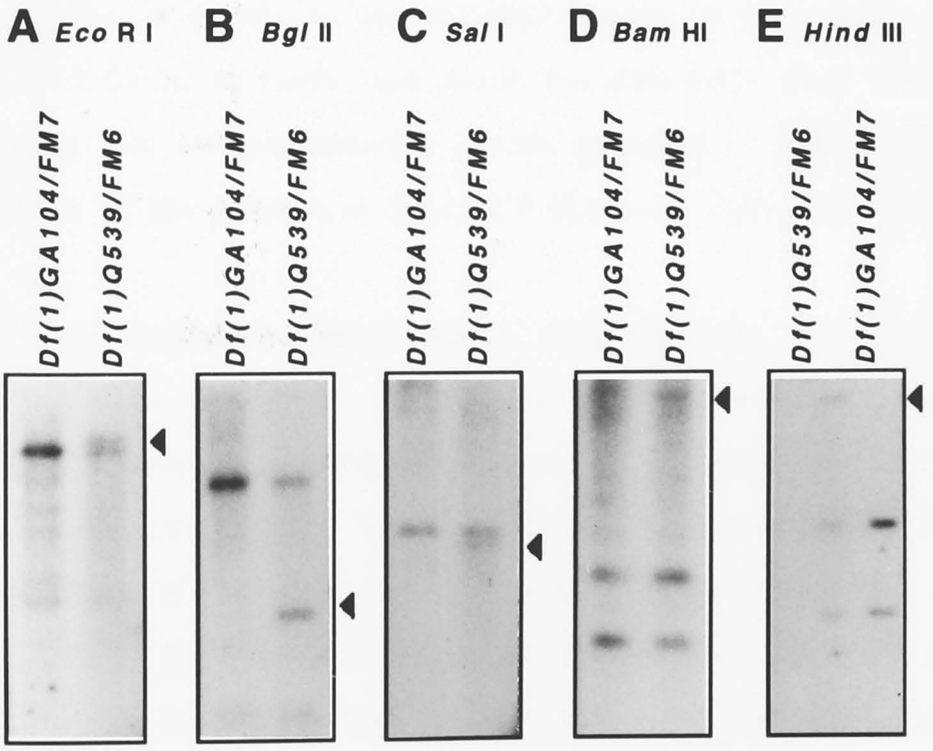


Fig. 2.8 Localization of the breakpoint in *Df(1)Q539* to position 1.7 to 6.7. Genomic DNA from the genotypes *Df(1)Q539/FM6* and *Df(1)GA104/FM7* was cleaved with the restriction endonucleases *Eco* RI (A), *Bgl* II (B), *Sal* I (C), *Bam* HI (D) and *Hind* III (E), separated by gel electrophoresis, blotted and hybridized with a labelled *Eco* RI/*Hind* III fragment (1.7 to 6.7 according to the co-ordinate system of Fig. 2.1). Novel fragments (11kb in A, 4kb in B, 5kb in C, 9kb in D and 14kb in E) corresponding to *Df(1)Q539* are found which are presumed to contain the deficiency.

Figure 2.8 illustrates the Southern transfers of genomic DNA from *Df(1)Q539/FM6* digested with 5 different restriction enzymes, in comparison to DNA from *Df(1)GA104/FM7* digested with the same enzymes. The breakpoint was mapped using the 1.5 to 6.5 *Eco* RI fragment as a hybridization probe. In each case a novel fragment, which is not present in the lane containing DNA made from the heterozygote *Df(1)GA104/FM7*, is observed in the lanes containing DNA made from the heterozygote *Df(1)Q539/FM6*.

2.3.4 Mapping of *Df(1)JC4*

Df(1)JC4 defines the proximal border of the *LB20* locus as *Df(1)JC4/LB20* is viable and hence the deficiency does not overlap the locus, but complementation groups proximal to *LB20* are within the region of the deletion in *Df(1)JC4* (Lefevre, 1981; Fig. 2.2).

An altered fragment which indicated the breakpoint was observed when DNA from position 23 to 39.5, cloned into a lambda phage and termed λ27-T8, was used as a radioactively labelled probe (Figs. 2.9, 2.10A). The hybridization pattern of *Df(1)JC4/FM7* was compared to that of *Df(1)B57/FM6* and to that of *Df(1)S54/FM6*. *Df(1)B57* has a breakpoint distal to *little fly-like*, and hence both chromosomes in *Df(1)B57/FM6* contain sequences corresponding to the loci *A112* and *LB20*. *Df(1)S54* uncovers the region from *uncoordinated* to *extra organs* including *A112* and *LB20*, hence sequences corresponding to these loci are only present on one chromosome in *Df(1)S54/FM6* .

Five *Hind* III fragments are visible at 4.2kb, 7.0kb, 8.2kb 10.5kb and 24kb, when λ27-T8 is used as a probe.

Figure 2.9

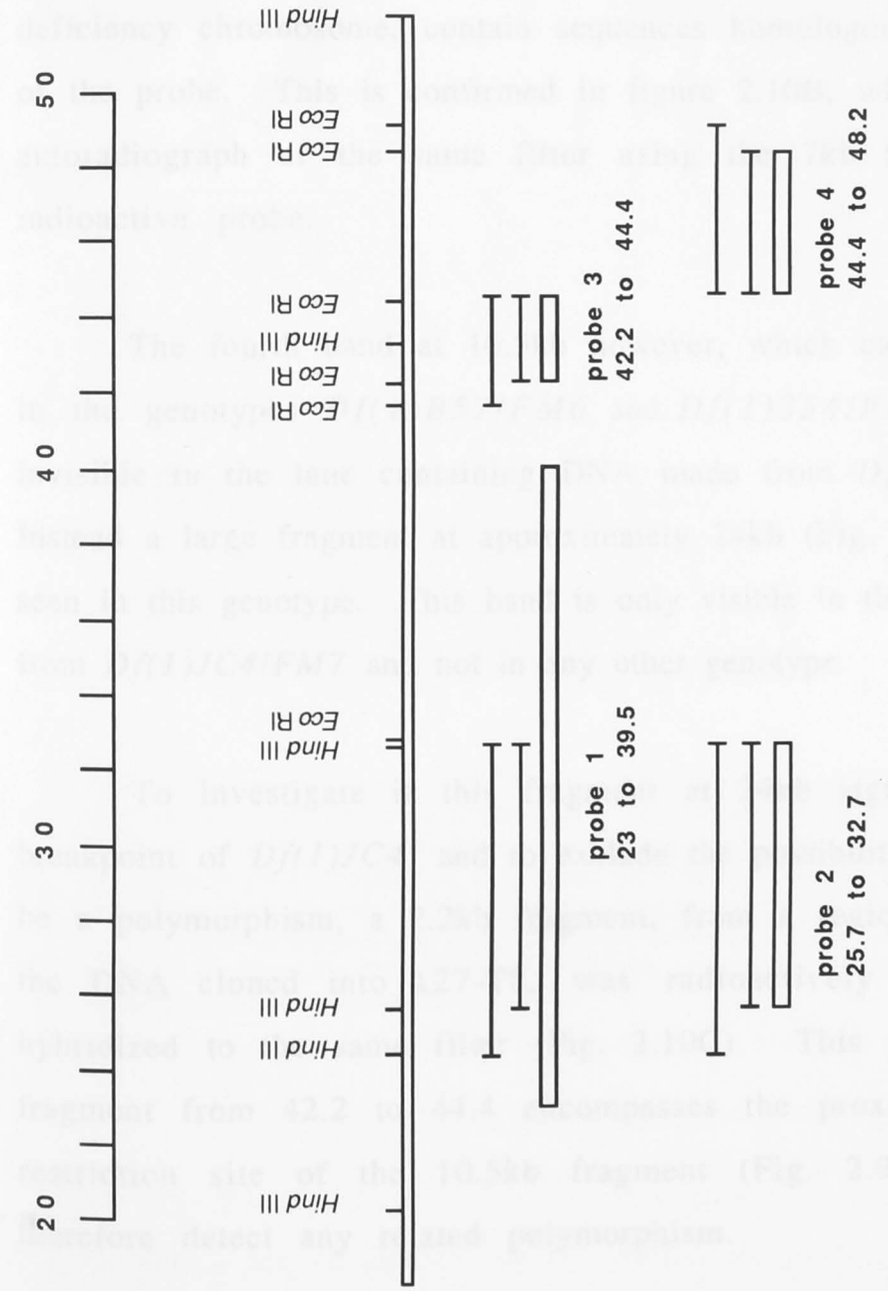


Fig. 2.9 Segment of the restriction enzyme map. The co-ordinate system above the restriction map is according to Fig. 2.1. The subclones used for the breakpoint analyses in Figs. 2.10 -2.12 are indicated below the restriction map as closed boxes. Observed polymorphic restriction sites are marked with bars corresponding to the two alternative fragments.

The *Hind* III restriction site at 25.5 (Fig. 2.9) appears to be polymorphic, since the signal at 7.0kb is visible in all genotypes tested, whereas the 8.2kb band does not appear in *Df(1)S54/FM6*, and the *Df(1)S54* chromosome has a deletion in this region. The fact that this polymorphism can be observed with *Df(1)JC4* allows the prediction that both chromosomes, the parental and the deficiency chromosome, contain sequences homologous to this part of the probe. This is confirmed in figure 2.10B, which shows the autoradiograph of the same filter using the 7kb fragment as a radioactive probe.

The fourth band at 10.5kb however, which can be observed in the genotypes *Df(1)B57/FM6* and *Df(1)S54/FM6*, is almost invisible in the lane containing DNA made from *Df(1)JC4/FM7*. Instead a large fragment at approximately 24kb (Fig. 2.10A) can be seen in this genotype. This band is only visible in the DNA sample from *Df(1)JC4/FM7* and not in any other genotype.

To investigate if this fragment at 24kb signals the distal breakpoint of *Df(1)JC4*, and to exclude the possibility that it might be a polymorphism, a 2.2kb fragment, from a region proximal to the DNA cloned into λ27-T8, was radioactively labelled and hybridized to the same filter (Fig. 2.10C). This 2.2kb *Eco* RI fragment from 42.2 to 44.4 encompasses the proximal *Hind* III restriction site of the 10.5kb fragment (Fig. 2.9) and should therefore detect any related polymorphism.

Sequences of the *Hind* III fragments hybridizing to the 2.2kb *Eco* RI fragment are visible as two bands at 10.5kb and

Note:
The faint band at 24kb is in the same region as the strongest marker band. However, the signal in lane one is unlikely to be spill over from the lane two as there is no signal at this position

in lane 3. It will be important to repeat this result.

Figure 2.10

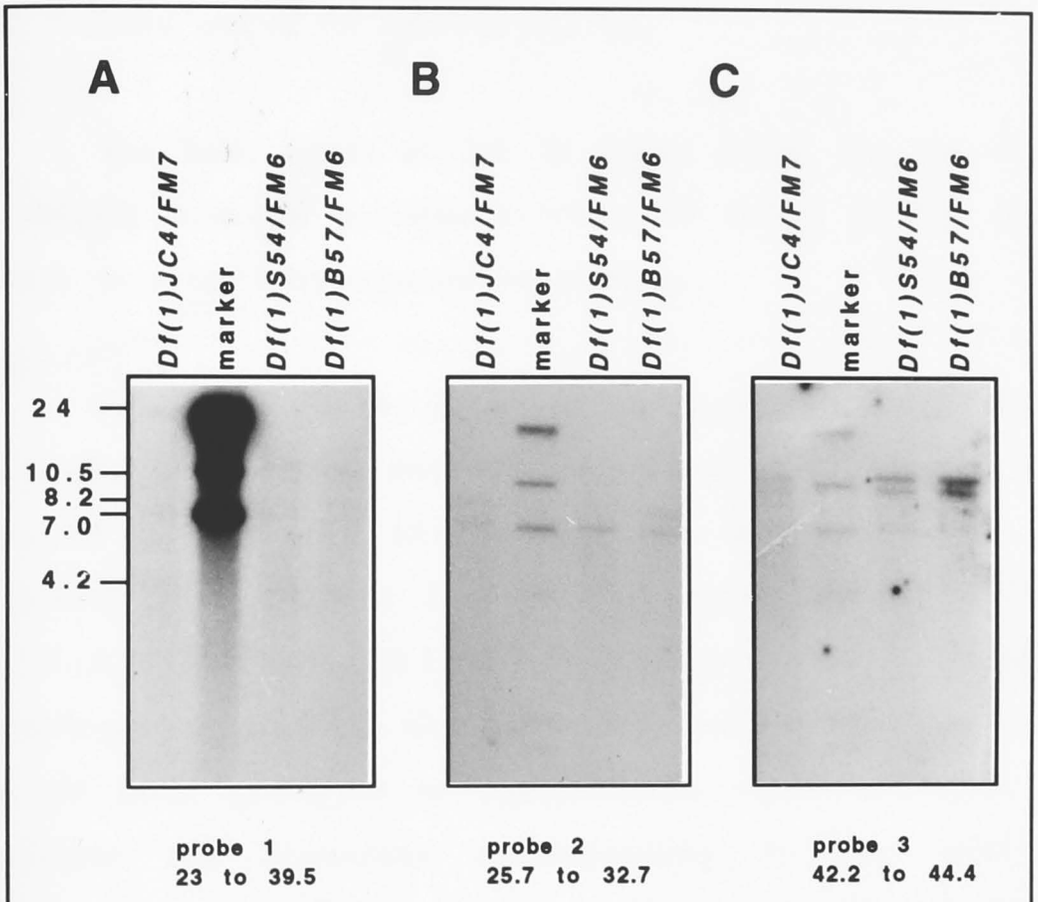


Fig. 2.10 Molecular mapping of the breakpoint in *Df(1)JC4*. Autoradiograms are presented of *Df(1)JC4* in comparison to *Df(1)S54* and *Df(1)B57*. Genomic DNA of the indicated genotypes was cleaved with *Hind* III, separated by gel electrophoresis, blotted and hybridized to labelled fragments. The probes used are indicated below each autoradiogram in terms of their position on the restriction enzyme map (according to the co-ordinate system in Fig. 2.1 and 2.9). Each probe is also shown in Fig. 2.9 with the same designation. The molecular sizes are indicated in kb.

9.2kb (Fig. 2.9, Fig. 2.10C). Two observations can be made. Firstly, both signals are of similar intensity within all genotypes and no signal is visible at 24kb. Secondly, the intensities of the signals in *Df(1)JC4/FM7* are much weaker then those in *Df(1)B57/FM6*. In comparison, the previous autoradiogram of the same filter (Fig.

2.10B) shows signals of similar intensities for these two genotypes. Thus, the distal breakpoint of *Df(1)JC4* has to be between the *Hind* III restriction site at position 32.7 and the position 39.5, which is the proximal end of the insert in λ27-T8.

The faint bands at 7kb in figure 2.10C are the signals remaining from the previous hybridization shown in Fig. 2.10B, which were not fully removed on washing.

To confirm that the 2.2kb *Eco* RI fragment from 42.2 to 44.4 is deleted in *Df(1)JC4*, this fragment, as well as the adjacent 3.8kb *Eco* RI fragment from 44.4 to 48.2, were hybridized to *Eco* RI digested DNA obtained from the genotypes, *Df(1)JC4/FM7*, *Df(1)B57/FM6* and *Df(1)S54/FM6* (Fig. 2.11A). Sequences homologous to the 2.2kb *Eco* RI fragment (42.2 to 44.4) are present in all three genotypes in approximately equal intensities and indicate the sequences corresponding to the wild-type chromosome. The lane containing DNA from *Df(1)B57/FM6* exhibits a second band at 3kb (see also Fig. 2.9), which is the polymorphic homologue to this fragment derived from the deficency bearing chromosome.

Similarly the sequences homologous to the proximal adjacent 3.8kb *Eco* RI fragment (44.4 to 48.2) have a polymorphic homologue at 4.3kb (Fig. 2.9, 2.11A). The sequences corresponding to this 4.3kb fragment derive from the wild-type chromosome. The lane containing DNA from *Df(1)B57/FM6* shows additionally the fragment at 3.8kb (Fig. 2.11A) which is the homologous fragment of the deficiency bearing chromosome.

To confirm that equal amounts of DNA were loaded onto this gel, the filter was also hybridized to a probe made from a fragment distal to the *Df(1)JC4* breakpoint (Fig. 2.11B). This probe was made

Figure 2.11

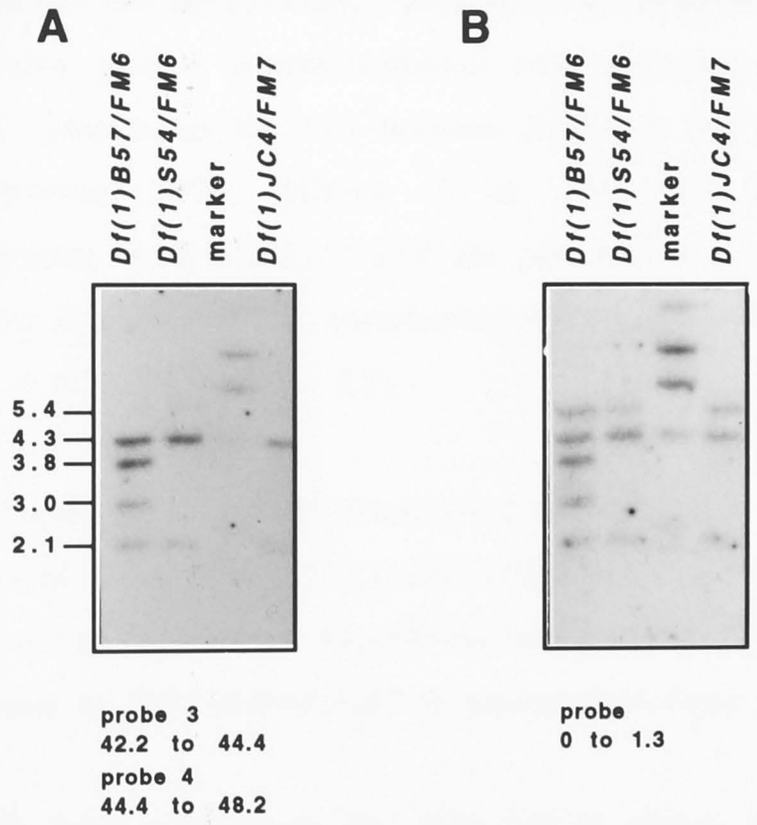


Fig. 2.11 Molecular mapping of *Df(1)JC4* (panels A and B) and (continued overleaf, in C to F, using the same probes)molecular mapping of *Df(1)GA104* and *Df(1)17-257* .
Panels A and B illustrate Southern transfers of the genotypes *Df(1)JC4/FM7*, *Df(1)S54/FM6* and *Df(1)B57/FM6*. The hybridisation with two sets of radioactive probes is presented. In panel A the transfers are hybridized with probe 3 and probe 4 of Fig. 2.9.
In panel B the same filter is hybridized with a radioactively labelled fragment from position 0 to 1.3. The relative intensity of hybridisation was known for this probe and was therefore used to confirm that equal amounts of DNA were loaded onto the gel.

from the 1.3kb *Hind* III fragment (position 0 to 1.3). The corresponding *Eco* RI fragment at 5.4kb (Fig. 2.11B) is, as expected, present in equal amounts in the genotypes D*f(1)B57/FM6* and *Df(1)JC4/FM7*.

2.3.5 Mapping of *Df(1)GA104* and *Df(1)17-257*

The molecular breakpoints of two additional deficiencies, *Df(1)GA104* and *Df(1)17-257*, have been investigated (Fig. 2.11C to F). Using genetic complementation tests deficiency *Df(1)GA104* fails to complement the loci between *flightless* and *LB20* (Schalet and Lefevre 1976, Miklos *et al.* 1987, see Fig. 2.2). In heterozygotes with *Df(1)17-257* the proximal loci *tumorous head* and *extra organs* are not complemented (Schalet and Lefevre 1976, Miklos *et al.* 1987, see Fig. 2.2).

Figure 2.11C and D illustrates a Southern blot comparing *Eco* RI digested DNA from *Df(1)GA104/FM7* and *Df(1)JC4/FM7*. The results of an equivalent experiment using *Df(1)17-257/Binsn* in comparison to *Df(1)HM44/FM7* is presented in figure 2.11 E and F.

In both experiments the same set of probes was used as in the previous experiment (Fig. 2.10), which are the two *Eco* R I fragments termed probe 3 and 4 (Fig. 2.11D,E). Also, that equal amounts of DNA were loaded onto the gel was again confirmed using the same probe as in figure 2.10, the 1.3 *Hind* III fragment which hybridizes to the band at 5.4kb (Fig. 2.11C,F) as was described above for the *Df(1)JC4* breakpoint (Fig. 2.11B). Only the wild-type chromosome is present in both deficiencies. The proximal breakpoints of these deficiencies,

Df(1)GA104 and *Df(1)17-257*, are therefore proximal to the breakpoint of *Df(1)JC4* and were not analysed further.

Figure 2.11

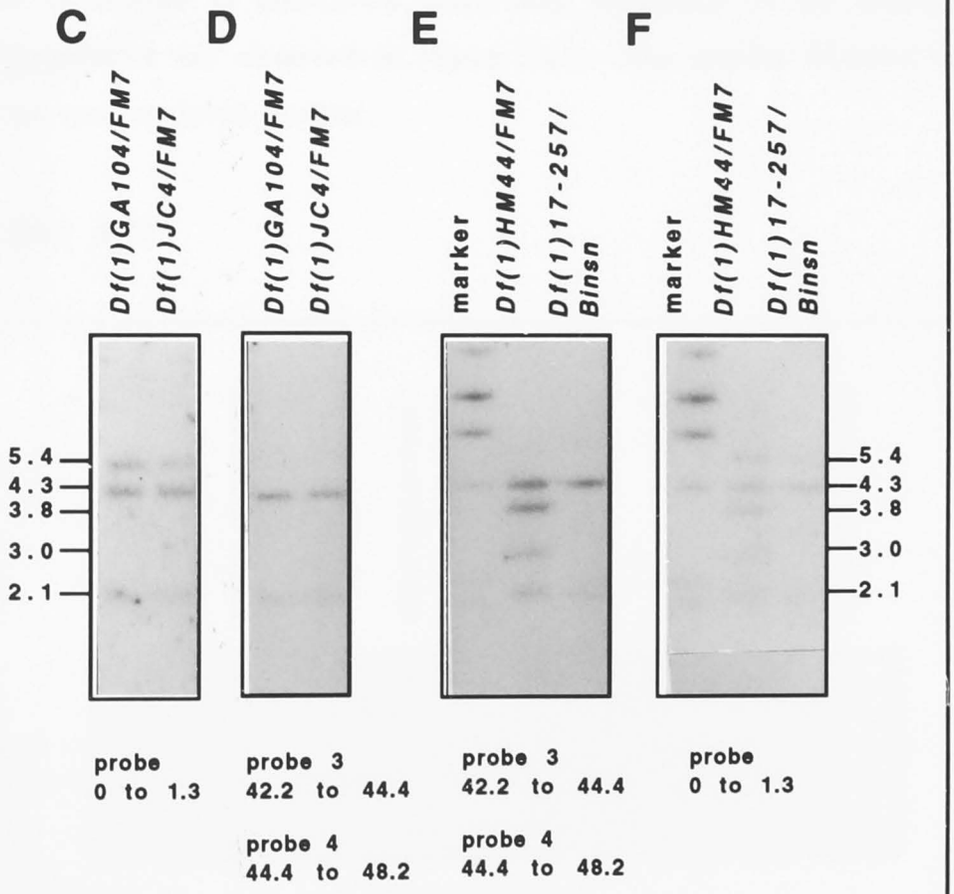


Fig. 2.11 continued
Molecular mapping of *Df(1)GA104* and *Df(1)17-257*.
Panels C and D illustrate Southern transfers of *Df(1)GA104/FM7* and *Df(1)JC4 /FM7*. Panels E and F illustrate Southern transfers of *Df(1)HM44/FM7* and *Df(1)17-257/Binsn* .
Equivalent to panels A and B, the hybridisation with two sets of radioactive probes is presented. In panels D and E the first hybridization is presented using probe 3 and probe 4 of Fig. 2.9. In panels C and F the hybridization of the same filters is illustrated using a distal fragment from position 0 to 1.3

2.3.6 Mapping of *Df(1)JC77*

The deletion in *Df(1)JC77* extends over the loci *A112*, *LB20*, *tumorous head* and *extra organs*. The molecular position of the distal breakpoint in *Df(1)JC77* has been mapped to the region -11.2 to -10 (David Hayward, personal communication). The molecular position of the proximal breakpoint was found to be proximal to the breakpoint of *Df(1)JC4* since only sequences of the wild-type chromosome are observed in figure 2.12. The precise location was thus not analysed further.

Figure 2.12

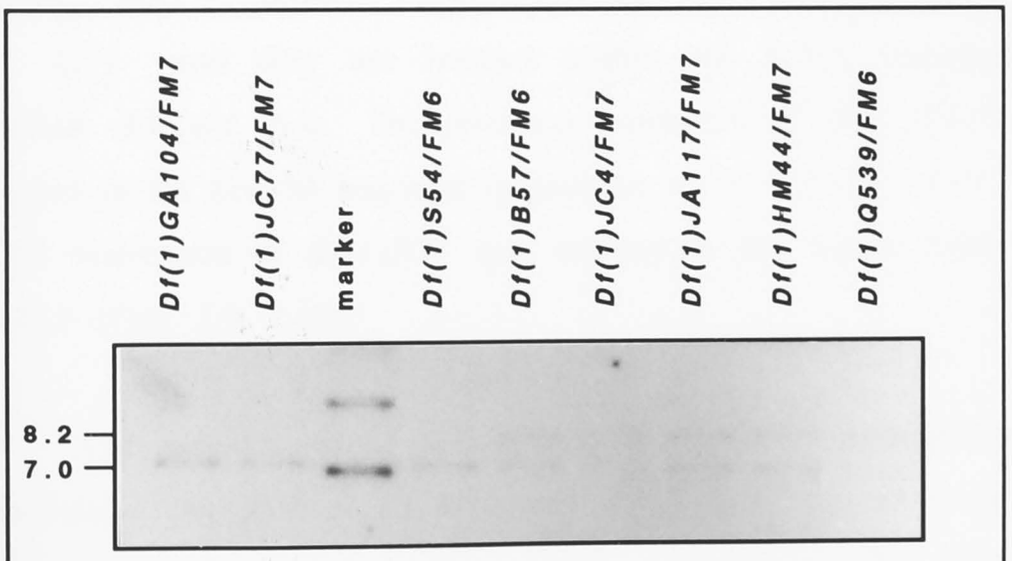


Fig. 2.12 The proximal breakpoint *Df(1)JC77*.
Genomic DNA from various genotypes (indicated above each panel) was cleaved with *Hind* III and separated by gel electrophoresis. After transfer, the filter was hybridized to a labelled *Hind* III fragment from position 25.7 to 32.7. The sizes of the two polymorphic fragments are indicated in kb. All those deficiency chromosomes known to contain the DNA of this *Hind* III fragment reveal two bands. The *Df(1)S54* chromosome, which is known to have this fragment deleted, shows one band, indicating that the *Hind* III restriction site at 25.7 is present in the parental but not in the mutant chromosome.
In the panel containing DNA made from *Df(1)JC77 /FM7* only one band at 7kb can be observed, suggesting that the proximal breakpoint of *Df(1)JC77* lies proximal to this fragment. The same argument holds for*Df(1)GA104* and *Df(1)S54*.

2.4 Discussion

With the intention of defining the genomic region occupied by the genes *A112* and *LB20*, the molecular breakpoints of a number of deficiencies were mapped.

The molecular positions of these deficiencies are summarized in figure 2.13, which shows the alignment of the cytogenetic and molecular maps. The proximal breakpoint of *Df(1)HM44* (Figs. 2.3, 2.4) was mapped to the *Eco* RI restriction site at -10. Both breakpoints of *Df(1)JA117* were mapped (Figs. 2.5, 2.6, 2.7) and they are located within the 2.2kb region from position -8.6 and -6.4. The proximal breakpoint of *Df(1)Q539* was mapped to the *Eco* RI fragment in position 1.5 to 6.5 (Fig. 2.8). The distal breakpoint of *Df(1)JC4* was mapped to the region from 32.7 to 39.5 (Figs. 2.9, 2.10).

The borders of a locus can be defined molecularly by establishing the position of deficiency breakpoints on the molecular map, using those deficiencies which complement the locus in question but not the closest known loci. The closest known locus distally to *LB20* is *A112* (Miklos *et al.*, 1986). Since *LB20/Df(1)HM44* is viable but *A112/Df(1)HM44* is lethal, *Df(1)HM44*, defines the distal breakpoint of the *LB20* locus. The closest known locus proximal to *LB20* is *tumorous head* (*tuh-1*) (Lefevre 1981). Because *Df(1)JC4* complements *LB20* but does not complement *tuh-1*, this deficiency was used to define the

Figure 2.13

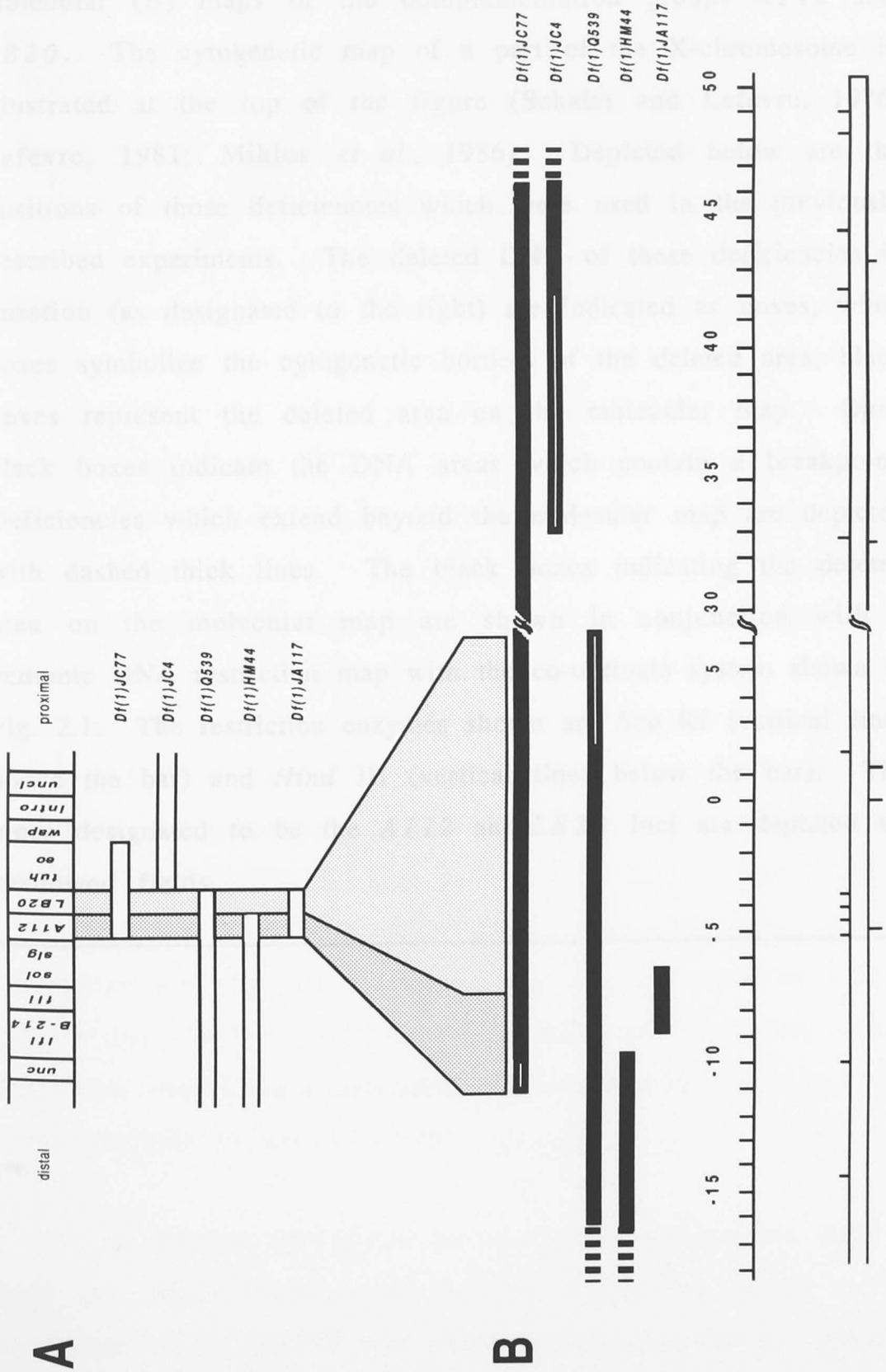

A
B
distal
proximal
unc
III B-214
III
sol alg
A112
LB20
tuh 60
wap
intro
uncl
Df(1)JC77
Df(1)JC4
Df(1)Q539
Df(1)HM44
Df(1)JA117
-15
-10
-5
0
5
30
35
40
45
50

Fig. 2.13 The area covered by the *LB20* and *A112* complementation groups. Alignment of the cytogenetic (A) and molecular (B) maps of the complementation groups *A112* and *LB20*. The cytogenetic map of a part of the X-chromosome is illustrated at the top of the figure (Schalet and Lefevre, 1976; Lefevre, 1981; Miklos *et al.*, 1986). Depicted below are the positions of those deficiencies which were used in the previously described experiments. The deleted DNA of these deficiencies in question (as designated to the right) are indicated as boxes, white boxes symbolize the cytogenetic borders of the deleted area, black boxes represent the deleted area on the molecular map. Open black boxes indicate the DNA areas which contain a breakpoint. Deficiencies which extend beyond the molecular map are depicted with dashed thick lines. The black boxes indicating the deleted area on the molecular map are shown in conjunction with a genomic DNA restriction map with the co-ordinate system shown in Fig. 2.1. The restriction enzymes shown are *Eco* RI (vertical lines above the bar) and *Hind* III (vertical lines below the bar). The areas designated to be the *A112* and *LB20* loci are depicted as shadowed fields.

proximal border of the *LB20* locus. The position of these breakpoints on the molecular map determines the DNA region which is covered by the *LB20* locus (Fig. 2.13).

The breakpoint of *Df(1)HM44* was mapped to the 4.8kb genomic fragment at position -10 to -5.2, (Figs. 2.3, 2.13) close to its distal *Eco* RI restriction enzyme site. The distal breakpoint of *Df(1)JC4* was mapped to the DNA area from 32.7 to 39.5 (Fig. 2.10). Thus the extent of the *LB20* complementation group covers 50kb of DNA between the proximal breakpoint of *Df(1)HM44* at -10 and the distal breakpoint of *Df(1)JC4* at 39.5 (Fig. 2.13). In an attempt to further define this region additional deficiencies were mapped. The deficiencies *Df(1)JC77*, *Df(1)GA104* and *Df(1)17-257* were shown to have deletions which included the entire *LB20* region.

However, the breakpoints of two deficiencies, *Df(1)Q539* and *Df(1)JA117*, provide data with which it is possible to reduce the region of *LB20* to 15kb. The genotypes *Df(1)JQ539/LB20* and *Df(1)JA117/LB20* are lethal (Schalet and Lefevre, 1976; Lefevre, 1981). On the molecular map *Df(1)JA117* was shown to be a small deletion of 1.5kb (Fig. 2.5 and 2.6). Its breakpoints were mapped to a 2.2kb area between position -8.6 to -6.4. The most obvious explanation for why this deletion in *Df(1)JA117* affects the *LB20* gene is that it deletes at least parts of the transcription unit. Thus one would expect the transcription unit of the *LB20* gene to be in close proximity to this breakpoint.

The deletion *Df(1)Q539* includes *LB20* and extends distally from this locus (Schalet and Lefevre, 1976; Fig. 2.13). The breakpoint of *Df(1)Q539* was placed between position 1.5 and 6.5 (Fig. 2.8). Therefore the conclusion is that the *LB20* gene probably

resides somewhere between this breakpoint at position 1.5 and 6.5 and the breakpoint of *Df(1)HM44*, at position -10, although formally the possibility remains that part or all of the *LB20* gene lies proximal to the breakpoint of *Df(1)Q539*.

Apart from a deletion directly affecting part of the *LB20* transcription unit, inactivation of the gene could also result from the new position into which the gene is placed due to the deletion. Euchromatic position effects due to neighbouring DNA sequences have been shown to affect the expression of transformed genes (Dutton and Chovnick, 1991).

Df(1)16-129 was used to define formally the distal border of *A112* (Schalet and Lefevre, 1976). The closest known loci distally to *A112* are *small optic lobes (sol)* and *sluggish (slg)* (Fischbach and Heisenberg, 1981; Miklos *et al.*, 1987). However, no deficiencies were available to separate *A112* from these loci. The proximal border of *A112* is, like *LB20*, therefore defined by *Df(1)JC4*, because no deficiency is available which is viable with *A112* and lethal with *LB20*.

Thus the *A112* locus is bordered by the proximal breakpoint of *Df(1)16-129* and the distal breakpoint of *Df(1)JC4* (Fig. 2.2). However, additional information can be used to reduce the size of this area, because this region also contains distally the genes *small optic lobes (sol), sluggish (slg),* and proximally *LB20*.

The *A112* gene has to be distal to the *LB20* gene, since *A112/Df(1)HM44* is lethal and *LB20/Df(1)HM44* is viable (Miklos *et al.*, 1986). Although at least part of the *A112* complementation group is within the deletion in *Df(1)HM44*, it is possible that parts

of the *A112* gene are proximal to the breakpoint at position -10 (Fig. 2.13).

The *A112* complementation group is also included within the deletion in *Df(1)JA117* which is a small deletion on the molecular map, of 1.5kb, within the region -8.6 to -6.4 (Figs. 2.5 and 2.6). Hence, the distal breakpoint of *Df(1)JA117* has to be between -8.6 and -7.9. The same argument as mentioned before for *LB20* applies also to *A112*, that is, the most likely reason why the deletion in *Df(1)JA117* affects the gene is that it physically deletes a vital part of the transcription unit. Thus one would expect the transcription unit of the *A112* gene to be in close proximity to the distal breakpoint of *Df(1)JA117* at -7.9 to -8.6.

Molecular data are also available for the complementation groups *small optic lobes (sol)* and *sluggish (slg)*, which are the loci distally adjacent to the *A112* locus. One of the deficiency breakpoints which defines *sluggish* is the breakpoint of *Df(1)JC77* (Lefevre, 1981; Miklos *et al.*, 1987; Fig. 2.13). The DNA deleted in *Df(1)JC77* extends from the locus *A112* proximally towards the complementation groups *tumorous head* and *extra organs*. The distal breakpoint of *Df(1)JC77* was mapped in the genomic area between position -10 and -11.2 (David Hayward, personal communication). Again, as mentioned previously, because the *A112* complementation group is within *Df(1)JC77*, it is possible that the *A112* gene extends distally across this breakpoint.

Molecular analyses of the area have shown that the most proximal transcript of the *sluggish* region impinges on the genomic *Eco* RI fragment at position -14.5 to -10 (David Hayward, personal communication). The simplest explanation is therefore that *A112*

lies in the region between the overlap of the deficiencies *Df(1)JC77*, *Df(1)JA117* and *Df(1)HM44* (Fig. 2.13), with the possibility that it might extend distally towards the *sluggish* transcript and proximally towards the *LB20* transcript.

This analysis of the results of the molecular mapping of the chromosomal deficiencies determines clearly the maximal borders for the complementation groups *A112* and *LB20* (Fig. 2.13). The sensitivity of the method is dependent on the availability of deficiencies in the vicinity of the genes in question. For example, the formally precise borders of the *A112* locus between the breakpoints of *Df(1)16-129* and *Df(1)JC4* are also the borders for three additional genes. However, in combination with additional information, it was possible to restrict the loci *A112* and *LB20* to a DNA region of approximately 16kb, which is now amenable to further analysis.

Two approaches for further analysis have been followed and are described in chapters 3 and 4. One is to analyse transcripts to match the genomic region with the transcription units. The second approach is an analysis at the nucleotide level to detect functional transcription signals in order to determine the borders of the genes.

Chapter 3

Analysis of transcripts in total RNA

3.1 Introduction

In eukaryotes the process of gene expression consists of four main stages, DNA transcription, post transcriptional processing, RNA translation and post-translational processing. Although gene expression has been observed to be regulated to a variety of extents at each of these stages, molecular analyses of a large number of genes indicate that differential gene expression is in most cases principally regulated at the level of messenger RNA (mRNA) production.

Because the regulation of transcription is one of the key stages at which gene expression can be controlled, the pattern of transcription is found to vary enormously between genes. This variation can be used to map transcription units within cloned genomic DNA by probing Northern blots with a series of contiguous genomic DNA subclones. By comparing the size and intensity of the hybridisation signals with the developmental profiles of the transcripts detected, a rough transcription unit map of a particular genomic DNA region can be constructed.

This approach will be described in this chapter. The transcriptional units of the region are investigated with the intention of comparing the location of these transcription units to the likely borders of the *A112* and *LB20* genes at the genomic DNA level.

The borders of the *A112* locus between the breakpoints of *Df(1)16-129* and *Df(1)JC4* include also distally to *A112* the genes *small optic lobes (sol) and sluggish (slg),* and proximally *LB20*. A detailed picture of the molecular position of these loci was established in the previous chapter, indicating that the *A112* locus maps to the overlap of several deficiencies around the position -10 on the molecular map (Fig. 2.13), extending distally towards the *sluggish* transcript and proximally towards the *LB20* transcript. The *LB20* locus maps proximally from position -10 to position 6.5 on the genomic map. In the following experiments this region is investigated at the transcriptional level.

3.2 Materials and Methods

3.2.1 Preparation of *Drosophila* RNA

This is a modification of methods described by Glisin *et al.* (1974) and Ulrich *et al.* (1977). Guanidinium thiocyanate was used to disrupt the cells, and the resulting homogenate was layered on a cushion of a dense solution of CsCl. Tissue was frozen in liquid nitrogen and ground to a fine powder using a cooled mortar and pestle. 8ml extraction solution (Extraction solution: 4M guanidinium thiocyanayte, 0.05M NaOAc, 1M β-mercaptoethanol, 0.005M EDTA) were mixed with charcoal and filtered through a 0.45μm Millipore syringe filter. Sarkosyl was added to a final concentration of 1% per gram tissue. The mixture was heated to 60°C until all solid materials were in solution before CsCl (0.5g/ml of extraction solution) was added. After all the CsCl was dissolved, 4ml of homogenate were carefully layered over 1.2ml of CsCl cushion solution (Cushion solution: 5.7M CsCl, 0.01M Tris-HCl pH7.5, 0.1M NaEDTA) in SW50.1 polyallomer tubes and spun at 35000 rpm for 16 hours at 15°C. During centrifugation, the RNA formed a pellet on the bottom of the tube, while the DNA and protein floated in the supernatant solution. The RNA was recovered from the gradient, dissolved in TE and precipitated with ethanol.

3.2.2 Analysis of RNA

Total RNA and mRNA were analysed using the Northern blotting technique (McMaster and Carmichael, 1977; Thomas, 1983).

(i) Electrophoresis of RNA

The procedures for electrophoresis of RNA were as described for DNA (see Section 2.2.9) with the following alterations. Up to 30µg total RNA or 5µg polyadenylated RNA was glyoxylated and separated on an agarose gel. Gels were poured and run in 10mM sodium phosphate buffer pH7.0 with constant recirculation of the buffer (1 litre of 500mM phosphate buffer contained 44.5g disodium hydrogen phosphate and 39g sodium dihydrogen phosphate). The RNA was glyoxylated by mixing 4µl Glyoxal, 10µl DMSO 2µl 100mM phosphate buffer and 5 µl RNA in a sterile Eppendorf tube and incubated at 50°C for 60 minutes (McMaster and Carmichael, 1977).

(ii) Transfer of RNA

RNA was transferred from agarose gels onto Hybond-N (Amersham) as described for DNA except that the denaturation and neutralization steps were omitted. After the transfer was completed, the filters were exposed to UV light for 5 minutes and baked for 2 hours at 80°C (after Thomas, 1983).

(iii) Hybridization of RNA to ^{32}P-labeled probes

The filter was prehybridized at 48°C for 1 to 2 hours in 25x phosphate buffer (250mM), 5x SSC, 7% SDS, 1mM EDTA and 50% deionized formamide. After the denatured radiolabelled probe was added directly to the prehybridization solution, the incubation was continued for 16 to 24 hours at 48°C. The filter was washed for 5 minutes at room temperature in 1x SSC, 0.1% SDS, followed by 3 washes of 20 minutes each at 68°C in 0.2x SSC, 0.1% SDS, after which the filter was wrapped in plastic film (Glad Wrap) and

exposed at -80°C to X-ray film (Kodak X-Omat XAR-5) using one intensifying screen.

To test whether equivalent amounts of RNA were bound to the Northern transfers, representative filters were hybridized to a radiolabelled probe made from the *D. melanogater ras* gene (kindly supplied by Dr David Hayward). This transcript is present in all developmental stages in approximately equal concentrations and was used to compare the quantity of the various RNA samples and to ensure that none of the RNA was degraded (Mozer *et al.*, 1985).

The RNA used to hybridize to probe 1 and 2 in Figure 3.1 was prepared by Dr. David Hayward. For all other analyses I prepared the RNA.

3.3 Results

To investigate the transcriptional activity of the two loci *A112* and *LB20*, total cellular RNA was isolated from different developmental stages. Samples of embryonic, combined larval, pupal and adult stages from the wildtype Canton S strain were examined for transcripts.

The probes were chosen so that the entire region was analysed for transcriptional activity. This region extended from the distal border of the *A112* gene, marked by the *sluggish* transcript, to the proximal border of the *LB20* gene, marked by the *Df(1)JC4* breakpoint. The region from -13 to 2 was analysed by probing with overlapping genomic fragments (Fig. 3.1). From the region 2 to 41 total RNA was analysed by hybrization with contiguous fragments (Fig. 3.2). Overall in the total region, four transcripts were observed (Figs. 3.1, 3.2, 3.3), two of these transcripts can be detected with probes from the -15 to 2 region (Fig.3.1) and two with probes proximal to this region (Fig. 3.2).

A small transcript of 0.6 kb was visible in larval and adult samples when a radioactively labelled DNA fragment corresponding to position 0 to 1.3 (probe 1 Fig. 3.1) was hybridized to total RNA. This 0.6 kb transcript was much more abundant in adult tissue than in larval tissue. A similar pattern of expression was also observed for a 0.6 kb transcript which hybridized to a probe from position -3.6 to 2 (probe 2 Fig. 3.1). It is most likely

Figure 3.1

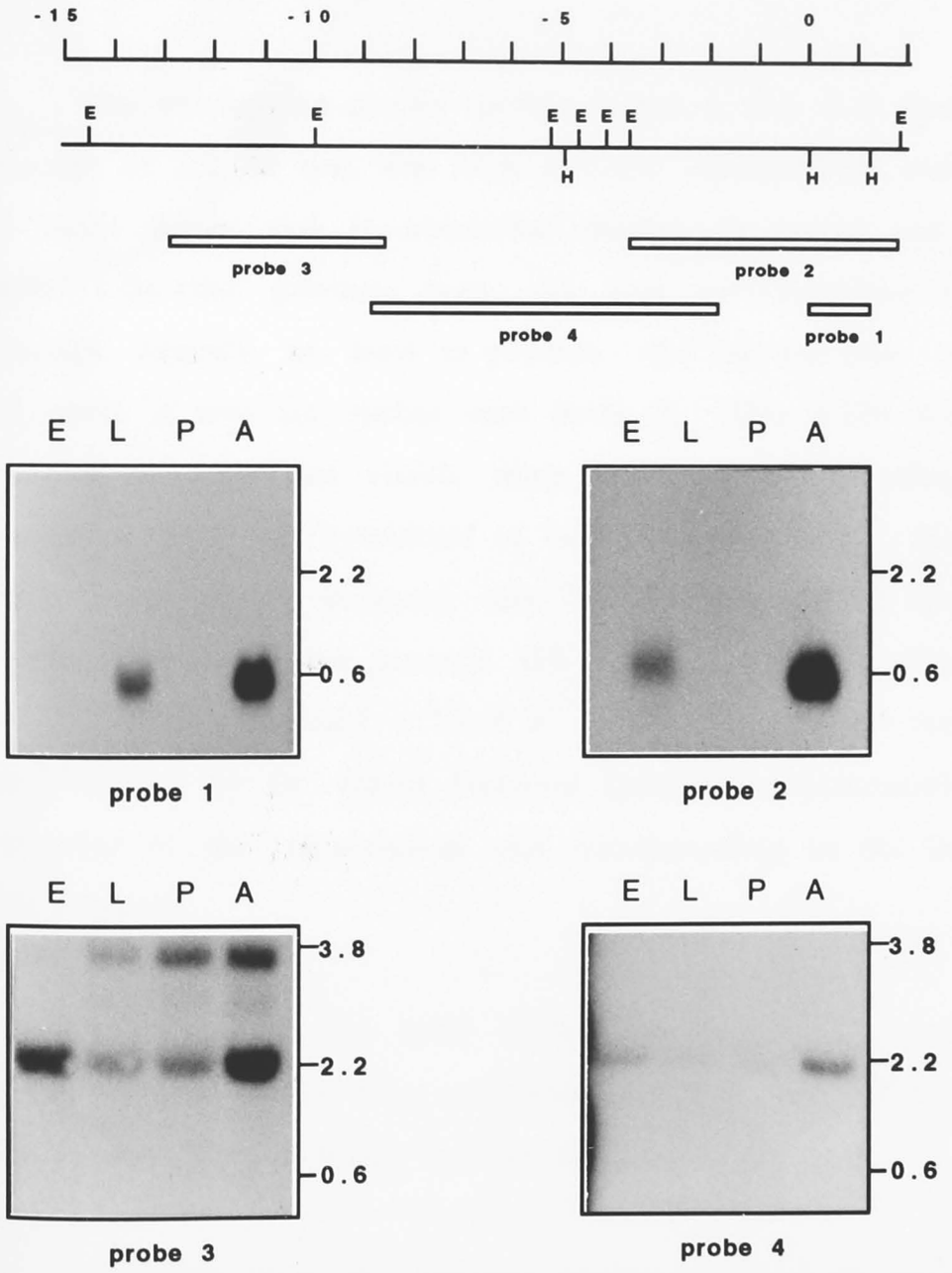


Fig. 3.1 Transcripts expressed in Canton S flies. Each lane contains 10µg of total RNA isolated from embryonic (E), combined larval stages (L), pupal (P) and adult (A) populations. The size of each transcript in kb is indicated to the right. Genomic fragments used as probes are shown in the upper part of the figure in conjunction with the genomic DNA restriction map (the co-ordinate system is according to figure 2.1). The restriction sites presented are *Eco* R1 (E) and *Hind* III (H). All autoradiograms were exposed for 10 days.

that both probes detect the same transcript because probe 2 overlapped probe 1 and because the size and the developmental profiles of the transcripts using these two probes appeared to be similar. No other transcript, except the 0.6kb transcript, was visible using these two probes.

Two overlapping probes (probes 3 and 4, Fig. 3.1) detected a transcript of 2.2 kb that was most strongly expressed in embryonic and adult tissue and is somewhat weaker in larval and pupal tissue. On the genomic map, the area corresponding to this transcript extends at least to position -5.2 but not past position -3.5, since it was not visible with probe 2. This 2.2kb transcript was also the only one visible when a 7kb probe extending from position -9 to -2 was hybridized to total RNA (Fig. 3.1). The most distal probe which detected this 2.2kb transcript is probe 3 (position -13 to -9), the genomic 4kb *Bam* HI fragment (Fig. 3.1). This probe also detected a band at 3.8kb. This 3.8kb transcript was identified by Dr. David Hayward (personal communication) as a member of the transcription unit corresponding to the adjacent locus *sluggish*.

Thus, probing the total RNA with overlapping fragments which cover the region from positions -13 to 2, two RNA species were detected, one of 0.6kb and one of 2.2kb.

Fragments from the genomic region proximal to position 2 were also used as probes to hybridize total RNA. Three probes were used to cover the region from position 2 to position 24 (Fig. 3.2). A transcript of 0.4 kb was observed with two

Figure 3.2

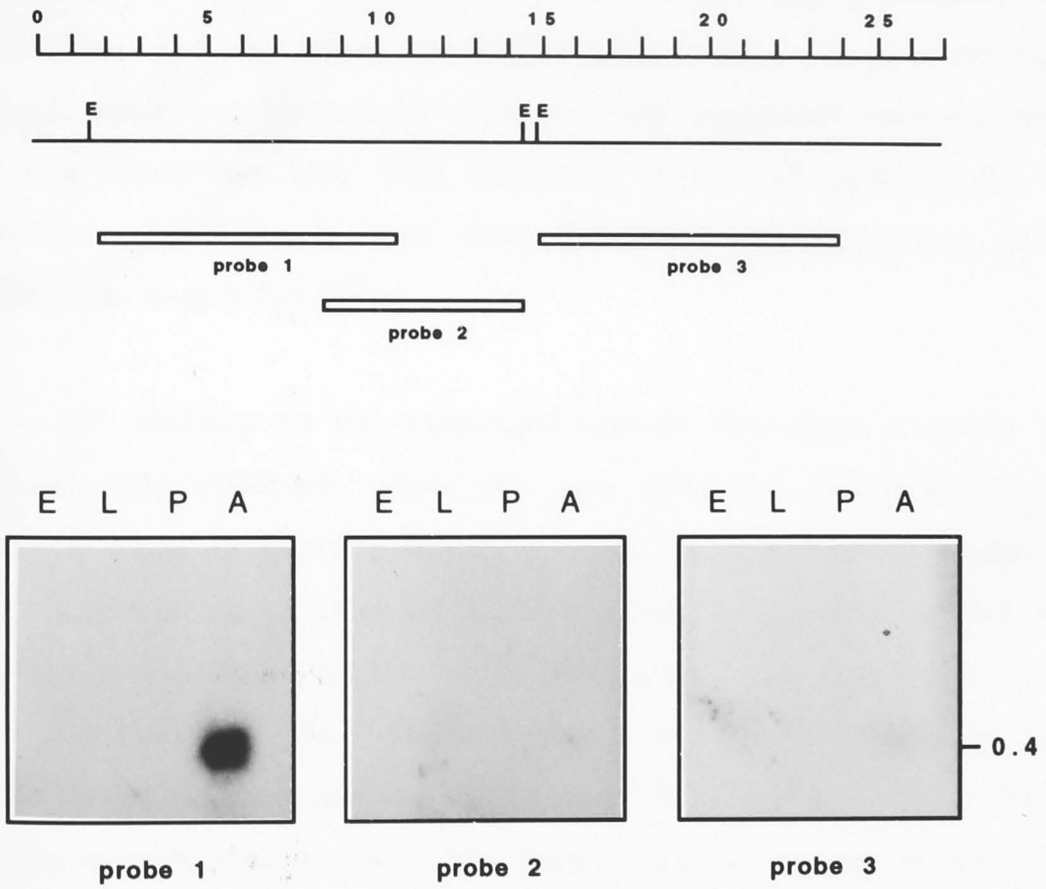


Fig. 3.2 Transcripts expressed in Canton S flies. Each lane contains 10 µg of total RNA isolated from embryonic (E), combined larval (L), pupal (P) and adult (A) populations. Both transcripts detected are 0.4 kb as indicated to the right. Genomic fragments used as probes are shown in the upper part of the figure which presents a segment of the genomic DNA restriction map with the co-ordinate system according to Fig. 2.1. Only *Eco* RI restriction sites are shown (vertical bars). The autoradiograms corresponding to probe 1 and 3 were exposed overnight, that corresponding to probe 2 was exposed for 14 days.

probes from this region (position 2 to 24) (Fig. 3.2). In each case the transcript was detected in adult tissue only. One of the probes corresponded to the genomic fragment from position 2 to 10. The other probe corresponded to the genomic fragment from position

15 to 24. The hybridisation signal to probe 3 appears very weak in the photography but was clearly discernible on the gel. This experiment needs to be repeated. These probes do not overlap and are separated by a 5.8 kb fragment (8.5 to 14.3) which fails to detect any transcript. Technically it is possible that a transcript of 0.4kb has had an intron of 5.8kb spliced out. A hybridization signal would not be present with genomic sequences corresponding to this intron but only with sequences distal and proximal to it. Another possibility is that there are two separate genes, both giving rise to a 0.4kb RNA.

In addition to the transcripts already described, a smear of signals was observed when the two proximal fragments from position 25 to 33.5 (probe 1, Fig. 3.3) and from 33.7 to 41 (probe 2, Fig. 3.3) were used as probes. This smear is probably caused by repetitive sequences within those fragments. In agreement with this hypothesis is the low background of signals visible in the autoradiograph shown in figure 2.10 (chapter 2). This autoradiograph also shows a low background of signals, in addition to the expected bands, when probe 1 is hybridized to genomic DNA. It is possible that some of the sequences in probe 1 are repeated elsewhere in the genome, therefore causing this background hybridization. A similar result was obtained when probe 2 was used on genomic DNA (data not shown). To exclude the possibility that this smear of signals is an artefact - perhaps due to degradation of the RNA - the filter hybridized with probe 2 was washed and hybridized to a probe made from the *ras* gene (Mozer *et al.*, 1985). No obvious RNA degradation could be detected in the autoradiograph (Fig 3.3).

Examiners please note that in figure 3.3 the autoradiograms corresponding to probe 2 and the ras gene have been transposed.

Figure 3.3

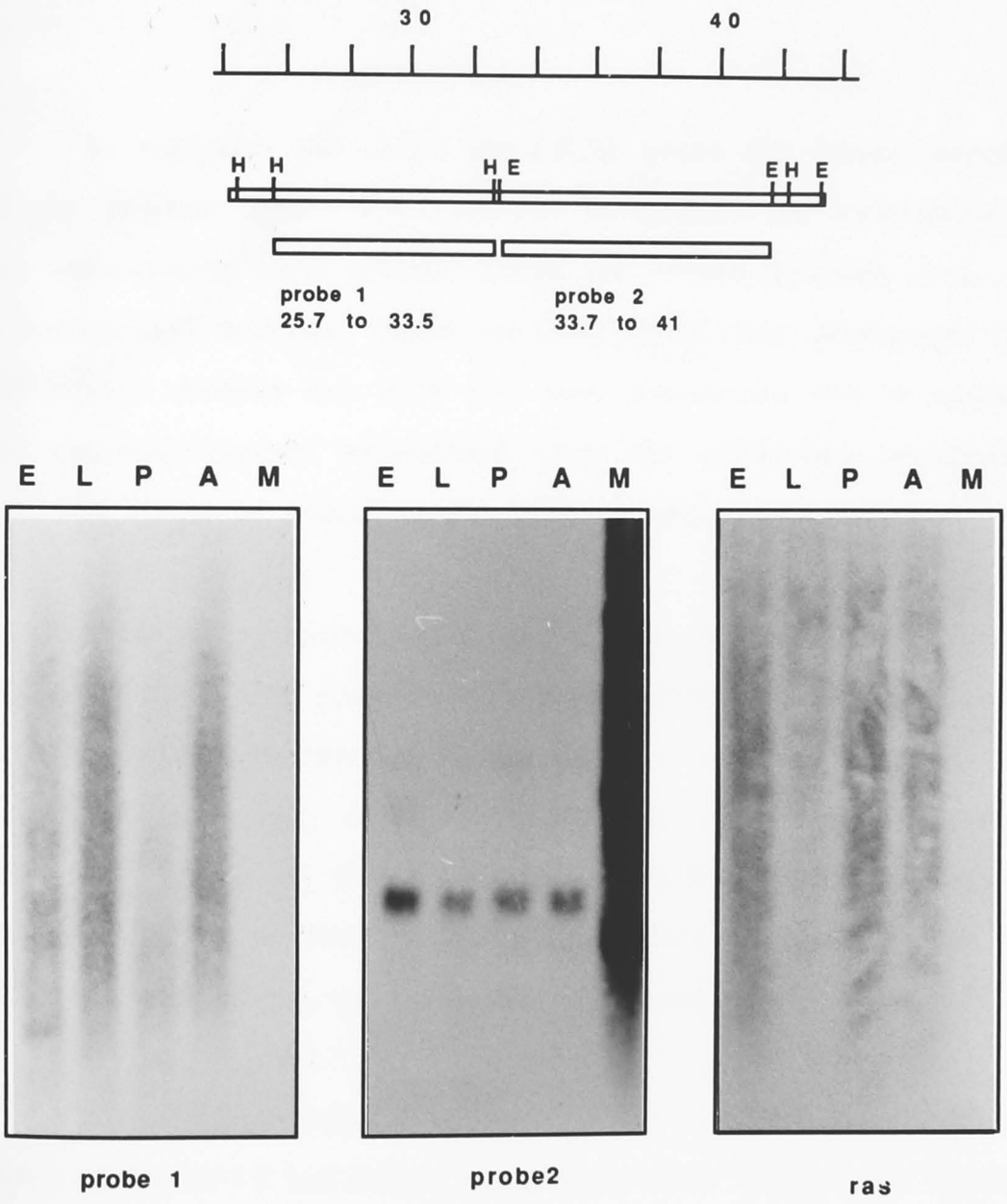


Fig. 3.3 Transcripts expressed in Canton S flies. Each lane contains 10μg of total RNA isolated from embryonic (E), combined larval (L), pupal (P) and adult (A) populations. Genomic fragments used as probes are shown in the upper part of the figure which presents a segment of the genomic DNA restriction map with the coordinate system according to figure 2.1. *Eco* R1 (E) and *Hind* III (H) restriction sites are shown (vertical bars). The autoradiograms corresponding to probe 1 and 2 were exposed for 5 days, that corresponding to probe 3, the *ras* gene, was exposed overnight.

3.4 Discussion

In analysing the *A112* and *LB20* genes the initial mapping of the genomic limits was extended to include the localization of the transcription units present within the cloned genomic area. A transcriptional analysis, based on total RNA, was performed with the aim to confirm that both loci were transcribed and to compare the genomic borders established from the deficiency breakpoints with the limits of possible transcription units.

Overlapping probes corresponding to the genomic DNA area from -10 to 2, and contiguous fragments from the region 2 to 41 were used to hybridize to total RNA in order to establish the transcriptional activity in the region. Thus the entire region from the distal border of the *A112* gene, marked by the *sluggish* transcript, to the proximal border of the *LB20* gene, marked by the *Df(1)JC4* breakpoint, was analysed for transcriptional activity.

Since the deletion mapping led to the hypothesis that the genes of the *A112* and *LB20* complementation groups are located distal to position 6.5, the emphasis of the analysis was on that DNA region. From position -13 to 2, smaller genomic fragments were used to pin-point more precisely the origin of possible transcription units.

Three RNA transcripts were detected in total RNA within the region -13 to 2. Firstly, a 2.2kb transcript corresponding to the DNA area from -13 to -6.8, secondly a 0.6kb transcript

corresponding to the DNA position 0 to 1.3 and thirdly a 0.4kb transcript corresponding to the DNA area from 2 to 6.5. The size, intensity and developmental profiles of the hybridisation signals of these transcripts are quite distinct and clearly describe three separate transcription units.

Deficiencies which define the likely borders of the A112 gene were mapped to the genomic region around position -10. These deficiencies were *Df(1)JC77*, which has a breakpoint within the fragment from -11.2 to -10 (David Hayward personal communication), *Df(1)HM44*, which has a breakpoint at -10 and *Df(1)JA117*, which has a 1.5kb deletion within the fragment from -8.6 to -6.4 (Fig. 2.13).

Genomic fragments in this region from -13 to -6.8 when hybridized to total RNA reveal a transcript of 2.2kb. Because of its position this transcript is deleted by *Df(1)JC77* and is also interrupted by the molecular breakpoint of *Df(1)HM44* and the deletion in *Df(1)JA117*. A probe from -13 to -10 which hybridizes to this 2.2kb transcript hybridizes also to a 3.8kb transcript identified by Dr. David Hayward as the most proximal transcript of the *sluggish* region (personal communication). Hence, the 2.2kb RNA species is the proximally adjacent transcript to the *sluggish* transcription unit. This transcript is therefore the most likely candidate for the *A112* transcription unit or represents at least part of it.

The *LB20* locus was localised to the DNA area from position -10 to 6.5, based on the molecular mapping of chromosomal

deletions (Fig. 2.13). This region is separated from the *A112* locus by the breakpoint of *Df(1)HM44*. Because *A112/Df(1)HM44* is lethal but *LB20/Df(1)HM44* is viable, *A112* and *LB20* are clearly separate and the *LB20* locus is placed proximally to the *A112* locus (Miklos *et al.*, 1986). The breakpoint of this deficiency also provides clear evidence that the 2.2kb transcript corresponding to sequences from -13 to -6.8 cannot be the *LB20* transcription unit, since it spans the breakpoint of *Df(1)HM44* at position -10.

The remaining two transcripts of 0.6kb and 0.4kb are possible candidates for the *LB20* transcription unit, since they map proximal to the breakpoint of *Df(1)HM44* and distal to the breakpoint of *Df(1)Q539*. One of these transcripts corresponds to the genomic sequences from 0 to 1.3. This 0.6kb transcript is very abundant in adults but less abundant in larval tissue. The second transcript is visible with a probe from position 2 to 10. It is 0.4kb in size and only detectable in adult tissue.

Both transcripts are reasonable candidates to be part of the *LB20* transcription unit as neither is produced by *Df(1)Q539*. Neither of these transcripts, however, is interrupted by the deletion in *Df(1)JA117*. Because one would expect to find the *LB20* transcript in close proximity to the proximal breakpoint of *Df(1)JA117*, the 0.6kb transcript appeared to be a more likely candidate than the 0.4kb transcript.

It is interesting that the probe corresponding to the 0.6kb transcript - the 0 to 1.3kb *Hind* III fragment - originally hybridized to a *Drosophila* clone, termed *DCg2*, which had been

isolated on its similarity to a chicken collagen probe (Natzle *et al.*, 1982). Based on this similarity of the *Hind* III fragment the corresponding 0.6kb transcript was tentatively described as "collagen-like" gene.

Chapter 4

Sequencing of the putative *A112* transcript

4.1 Introduction

The results described in chapters 2 and 3 suggest that *A112* resides in the region of the overlap between the deficiencies *Df(1)JC77* and *Df(1)HM44* (Fig. 2.13), but may extend distally towards the *sluggish* transcript and proximally towards the *LB20* transcript. A single transcript of 2.2kb was observed within the area of -11 to -6.8. This transcript is within the deletion of *Df(1)JC77* and is also interrupted by the molecular breakpoint of *Df(1)HM44*. It therefore belongs most likely to the *A112* transcription unit or at least represents part of it.

It is tempting to suggest that this transcript is the corresponding message to the *A112* gene. Two approaches for further analyses are based on these results. One approach is to determine if this transcript is indeed expressed from a gene and, if so, what is the protein coded. A second approach is to test whether this transcript actually belongs to the *A112* gene. The most conclusive test for the latter is to reintroduce the DNA in question into the genome and analyse whether it is able to complement mutations in the gene. This chapter is concerned with the nucleotide sequence of a region of the *A112* transcription unit and the transformation experiments are described in chapter 6.

The invention of rapid sequencing methods (Sanger *et al.*, 1977; Maxam and Gilbert, 1977) has greatly increased our understanding of eukaryotic gene structure and expression. The comparison of nucleotide sequences has developed into an

important tool to establish conserved functional motifs, for example those involved in the regulation of transcription, such as TATA boxes (Corden *et al.*, 1980; reviewed in Mitchell and Tijan 1989), intron-exon junctions (Mount, 1982), poly-adenylation sequences (Proudfoot and Brownlee, 1976; reviewed by Manley 1988) and those sequences which play an important role in development such as the homeobox (McGinnis *et al.*, 1984a; Scott and Weiner, 1984; Ingham, 1988; Akam, 1989).

Numerous genes important for development in vertebrates have been identified from their sequence similarity to *D. melanogaster* regulatory genes (Gaunt *et al.*, 1988; Duboule and Dolle, 1989; Graham, *et al.*, 1989; reviewed in Kessel and Gruss, 1990). The identification of functional relationships between seemingly unrelated genes, like the mouse mammary oncogene *int-1* which is homologous to the *D. melanogaster* segment polarity gene *wingless*, (Rijsewijk *et al.*, 1987), or the *Rel* oncoprotein (reviewed in Gilmore, 1991) which is similar to the *D. melanogaster* gene *dorsal* (Steward, 1987) and the transcription factor NF-κB (Gosh *et al.*, 1990; Kiernan *et al.*, 1990; Ruben *et al.*, 1991), were direct results from the comparison of sequenced genes.

As sequencing data accumulated another approach towards identifying genes developed using the knowledge of motifs which were conserved and hence might be structurally important. Examples include the recently defined RNA helicase family established from the basis of a few common motifs (Linder *et al.*, 1989), the POU family which was defined by the sequence homology of three mammalian transcription factors and one nematode regulatory protein (Herr 1988) and the TEA domain,

which is a new DNA binding motif (Buerglin 1991) discovered using computer-assisted searches of databanks.

The sequence analysis of the 2.2kb transcript, from -11 to -6.8, was therefore a logical extension of the mapping experiments described in the previous chapters. The first aim was to investigate whether a corresponding cDNA exists and, if so, what are its properties at the nucleotide level. The second aim was to assess the structural organisation of the primary sequence, to determine the number of introns and the extent of the transcription unit. It was also possible that knowledge of the sequence might give clues to the function of the gene, by comparing its sequence to that of other sequences in the nucleotide or protein data bases.

4.2 Materials and Methods

4.2.1 Description of the cDNA libraries

A library in bacteriophage λgt^{10} (Poole *et al.*, 1985) was screened for cDNAs. This cDNA library was a combination of three libraries made from 0-3 hr, 3-12 hr and 12-24 hr embryos. The following is a brief description of the protocol used by the authors (Poole *et al.*, 1985) to prepare these libraries. The poly(A)$^+$ RNA was primed with oligo(dT) and the first strand of the cDNA was made with reverse transcriptase. RNA was digested with RNAse A and the cDNA was separated from primer, deoxy- and ribonucleotides on Sepharose CL-2B. Tailing of the first strand with dG was accomplished using terminal transferase. Oligo(dC) was annealed to the tailed first strand and the second strand was synthesized using Klenow fragment of DNA polymerase I. Internal *Eco* RI sites were methylated, *Eco* RI linkers were ligated and after cutting with *Eco* RI the cDNA was ligated into bacteriophage λgt^{10}. The bacteriophage were packaged *in vitro*. The authors claim that 50% of the clones had inserts larger than 750bp and that they were able to obtain several clones of 2.5 to 2.6kb in length.

4.2.2 Library screening

The bacteriophage λgt^{10} were plated on *E. coli* strain JP777 and screened using standard procedures (Sambrook *et al.*, 1989). Lambda bacteriophage were screened with radioactive labelled probes. Lambda bacteriophage to be screened were plated to a density of 100-300 plaques per 9cm petri-dish. Circular nitrocellulose filters (Hybond-C extra, Amersham) were marked

with a marker pen and placed onto the surface of the plate and allowed to wet. The orientation marks were copied onto the petri-dishes. The filter was peeled off and the DNA attached to it was denatured for 10 minutes in 0.8M NaCl, 0.4M sodium hydroxide (NaOH), neutralized for 10 minutes in 1.5M NaCl, 0.5MTris pH8 and washed in 2xSSC (0.3M NaCl, 0.03M trisodium citrate). The filters were air-dried and baked for 2 hours at 80°C under vacuum. The plaques containing the DNA sequences of interest were identified by hybridization to a gel-purified radioactive labelled probe in the same manner as described for Southern blotting. Bacteriophage DNA was prepared as described in chapter 2.2.

4.2.3 Cloning of DNA for Sequencing

(i) Subcloning into *pEMBL*

The DNA to be sequenced was cloned either into *pEMBL* vector *mp8+* (Dente *et al.*, 1983; a gift from Dr. David Hayward and Dr. Steven Delaney) or into the M13 vectors *mp10*, *mp18* or *mp19* (obtained from Boehringer Mannheim). Subcloning of individual DNA fragments into *pEMBL8+* and the M13 vectors *mp18* and *mp19* was carried out using gel-purified fragments (see Section 2.2.7) which were ligated with vector DNA.

DNA-fragments with cohesive ends were ligated in 50mM Tris-buffer pH7.6, 10mM $MgCl_2$, 1mM DTT, 5mM adenosine-5'-triphosphate (ATP) and 1mM spermidine with 0.1 Weiss unit of T_4 ligase. Typically vector and target DNA were ligated together in a molar ratio of 1:3 with a total amount of DNA of 0.2µg in a total volume of 10 µl at 4°C overnight.

Blunt end ligations were performed as described above with the following alterations: the ATP concentration was decreased to 0.5mM, the amount of T_4 ligase was increased to 4 Weiss units and

no spermidine was added. Approximately 100ng of ligated DNA was used to transform JM101 cells which had been made competent using the calcium chloride-method as described in Maniatis *et al.*, (1982).

(ii) Subcloning of randomly overlapping DNA into M13 *mp10*.

Sufficient amounts of DNA were cleaved with the appropriate restriction endonuclease to yield at least 10 µg of target DNA. The target DNA was purified using agarose gel electrophoresis and electroelution (see Section 2.2.6). The purified DNA was then self ligated and fragmented by sonication.

(iii) Size selection of DNA

The fractured DNA was then end repaired with the Klenow fragment of *E.coli* DNA polymerase I in a total volume of 25µl containing 50mM Tris-buffer pH7.5, 10mM $MgSO_4$, 1mM DTT and each of the four 2'-deoxynucleoside-5'-triphosphates at a concentration of 2mM. The reaction was incubated at room temperature for 30 minutes. The DNA was separated on an agarose gel in order to isolate DNA fragments of the appropriate size. DNA fragments from 250bp to 1200bp in length were excised, purified by electroelution (see Section 2.2.6) and ligated into M13 mp10. Approximately 400 ng purified DNA and 250 ng vector DNA were used per ligation. Competent cells (JM101) were transformed with these subclones.

(iv) Preparation of competent cells according to Hanahan (1983)

Cells were made competent following the method developed by Hanahan (1983). A scrape of bacterial cells (*E.coli* strain JM101) was taken from a frozen stock, streaked onto a SOB-plate (2.2.1)

and incubated at 37°C overnight. The following morning 3 colonies were picked to inoculate 20ml SOB-medium and shaken at 37°C until the optical density at 600nm reached 0.55-0.65. The cells were chilled on ice for 10-15 minutes and then harvested by a centrifugation at 4°C for 12 minutes at 3,500 rpm. The cells were drained and resuspended in a third of the original volume of transformation buffer (TFB: A 0.5M solution of 2N-morpholino-ethone sulfonic acid was equilibrated with potassium hydroxide to pH 6.3 and filter sterilized. To make up 1 litre of TFB, 20 ml of this solution was mixed with the following salts : 7.4g KCl, 8.9g manganese chloride, 1.5g $CaCl_2$ and 0.8g hexamincobalt (III) chloride. The volume was adjusted to 1 litre with double distilled water.). After incubating on ice for 10 - 15 minutes, the cells were pelleted as before and resuspended in one twelfth of the original volume of TFB. DnD solution was added to a final concentration of 3.5% (DnD: 100µl of 1M potassium acetate pH7.5 were added to 9 ml dimethyl sulfate (DMSO) and 1.53g DTT. The volume was adjusted to 10ml with double distilled water.). After 10 minutes incubating on ice a second aliquot of DnD was added to a final concentration of 7%. The cells were then competent and ready to be transformed.

To carry out the transformation an aliquot of DNA (typically one half of a ligation but not more than 20µl) was mixed with 200µl competent bacterial cells and incubated on ice for at least 40 minutes. The cells were given a heat shock by placing them in a waterbath at 42°C for 2 minutes, then chilled on ice and plated out. The amount of ligation mix used to transform JM101 cells which had been made competent by this method (Hanahan, 1983), was adjusted to yield not more than 250 plaques per plate. Successful recombinants were identified as colourless plaques.

4.2.4 Preparation of DNA for Sequencing

(i) Preparation of single stranded M13 DNA

Individual recombinant M13 clones were picked into 2 ml of LB medium and shaken over night at 37°C. The bacterial cells were removed by 5 minutes centrifugation in an Eppendorf centrifuge. The supernatant was transferred to a new Eppendorf tube and spun for a further 5 minutes to ensure that all the bacterial cells had been removed. The phage suspension was added to 250µl of a 20% polyethylene glycol (PEG), 2.5M NaCl solution, mixed and incubated for 15 minutes at room temperature. The single stranded M13 was precipitated by 5 minutes centrifugation. The entire supernatant was removed carefully using an aspirator and the pellet containing the single stranded DNA was dissolved in 100µl TE (pH8). The DNA was extracted with phenol and precipitated, the subsequent pellet was dissolved in 20 µl TE (pH8).

(ii) Preparation of single stranded *pEMBL* DNA

One day after transformation individual white colonies were used to inoculate 2ml of LB medium and grown by shaking at 37°C. After 1 hour the cells were superinfected with 1µl (10^6-10^7 plaque forming units) of single stranded helper phage M13K107 and shaken overnight at 37°C. The subsequent preparation of single stranded DNA was carried out as described for M13.

(iii) Preparation of double stranded *pEMBL* DNA

Double stranded *pEMBL8+* DNA for sequencing was prepared as described for large scale plasmid preparation (2.2.3).

4.2.5 Sequencing techniques

(i). Dideoxy Sequencing

The nucleotide sequences of the target DNA were determined by the dideoxy mediated chain-termination method (Sanger *et al.*, 1977). Sequencing reactions were carried out using [^{35}S]dATP αS and a sequencing kit (Sequenase Version 2.2), commercially supplied by United States Biochemical Corporation. Primers used in the annealing reaction were commercially supplied by United States Biochemical Corporation (universal and reverse primer) or by the Protein/DNA Facility of the Australian National University. The sequencing reactions were separated on denaturing polyacrylamide gels (7M urea, 5% polyacrylamide) cast in a Bio-Rad Sequigen apparatus. The samples were denatured at 80°C for 2 minutes just before loading onto the polyacrylamide gel. After electrophoresis at 110 Watts the gel was fixed in 12% methanol, 10% glacial acetic acid and dried on a slab gel drier. The gel was then exposed to an X-ray film (Kodak XAR-5), usually overnight at room temperature.

(ii) Sequencing of denatured double stranded DNA

Approximately 5μg double stranded DNA cloned into *pEMBL* was denatured with 4μl 1N NaOH in a total volume of 20μl and incubated at room temperature for 5 minutes. The sample was neutralized with 2μl of 5M ammonium acetate (pH4.6), immediately vortexed and precipitated with 50μl ethanol. After 15 minutes incubation on ice the denatured DNA was sedimented for 15 minutes in an Eppendorf centrifuge. The DNA pellet was dried and redissolved in 7μl double distilled water. Subsequent sequencing reactions were carried out in the same manner as those using single stranded DNA (according to the Sequenase Version 2.2 protocol).

4.2.6 Computer analysis

The sequence data generated by the "shotgun" strategy were assembled using BATIN, DBCOMP and DBUTIL programs of Staden (1984). The completed sequence was analysed with the SEQ program (Brutlag *et al.*, 1982) and MacVector (International Biotechnologies, Inc). The DATAbases NBRF protein bank (release 29.0), the Swiss protein bank (release 14.0) and the Genpept library (release 63.0) were searched for similar sequences using the FASTA algorithm (Pearson & Lipman, 1988). The searches were done using the Wisconsin Package (version 7.0, April 1991; copyright 1991 by John Devereux; Devereux *et al.*, 1984).

Ambiguous sequences were eliminated by sequencing both strands using dGTP (deoxyguanosine triphosphate) as well as dITP (deoxyinosine triphosphate). Also, independent subclones were sequenced until each nucleotide position was confirmed in at least six independent clones. On request these sequencing data can be accessed.

4.3 Results

4.3.1 Isolation of cDNA clones

In order to obtain more information about the *A112* transcript I decided to sequence cDNA clones derived from the region to which *A112* was localized. A cDNA termed pc37-3.8, corresponding to a genomic fragment from -10 to -13.5, previously isolated from an embryonic cDNA library (Poole *et al.*, 1985) and provided by Dr David Hayward, was cloned by the shotgun technique (Bankier and Barrell, 1984) and sequenced (Sanger *et al.*, 1977). To verify the genomic address of the pc37-3.8 cDNA, it was radioactively marked and hybridized to genomic DNAs, which had been cut with *Eco* RI (Fig 4.1). The two visible bands obtained were estimated to be 4.5kb and 4.8kb, as expected from the genomic restriction map (Fig. 4.1 and Fig 3.1, chapter 3).

It was later discovered when the genomic sequence was compared to the cDNA sequence that the pc37-3.8 cDNA was not a true reflection of the transcript corresponding to the *A112* region but was a rearranged cDNA, which probably occurred during the process of library construction.

Therefore a new set of cDNAs were isolated from a cDNA library made from adult flies (Poole *et al.* 1985) using two probes, the genomic 4.8kb *Eco* RI fragment from position -10 to -5.2 and

Figure 4.1

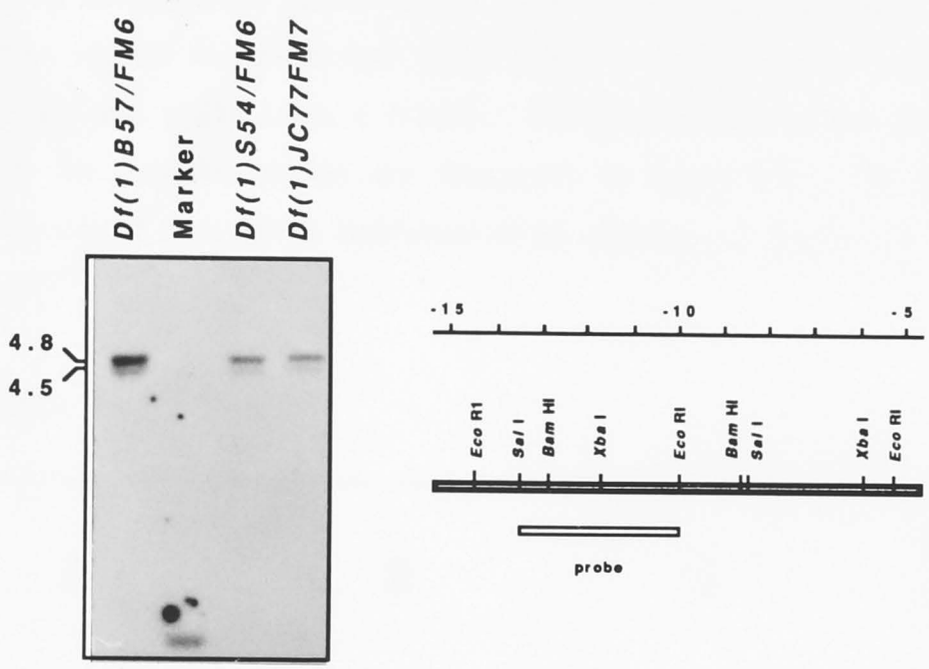


Fig. 4.1 Southern blot analysis of the *A112* cDNA pc37-3.8.
Southern transfers of genomic DNA of various genotypes (as indicated) which had been cut with *Eco* R1 were hybridized to the the *A112* cDNA pc37-3.8. Illustrated to the right is the restriction map, indicating the position of the genomic fragment used as a probe to isolate the *A112* cDNA pc37-3.8. The co-ordinate system is according to figure 2.1 in chapter2. Each panel shows two signals of 4.5kb and 4.8kb, corresponding to the genomic *Eco* R1 fragments from -13.5 to -10 and -10 to -5.2 respectively, indicating that the cDNA originates from this genomic area. The intensity of the hybridization signal is, as expected, stronger in *Df(1)B57/FM6* compared to *Df(1)S54/FM6* and *Df(1)JC77/FM7*, which is also in agreement with the hypothesis of a cDNA originating from the *A112* locus.

the adjacent 4.5kb *Eco* RI fragment from -14.5 to -10. One of those fragments, the 4.8kb *Eco* RI, fragment appears to be repeated elsewhere in the genome. Therefore screening with both probes was necessary in order to focus on the transcript of this particular locus. This second probe, the 4.5kb *Eco* RI fragment, encompasses the sequences which had previously been used to isolate the first set of cDNAs, including pc37-3.8 (Fig. 4.1).

The new cDNAs isolated in this second screen are termed pc5a, pc5b and pc21. All three contain an internal *Eco* RI site, which divides each cDNA into a small and a large fragment. The sizes of the fragments are: pc5a: 1.4kb + 0.55kb, pc5b: 1.4kb + 0.75kb and pc21: 1.0kb + 0.65kb. The relationships of the cDNAs and the genomic region are illustrated in figure 4.2. The large fragment of each cDNA hybridizes to the genomic 4.8 kb *Eco* RI

Figure 4.2

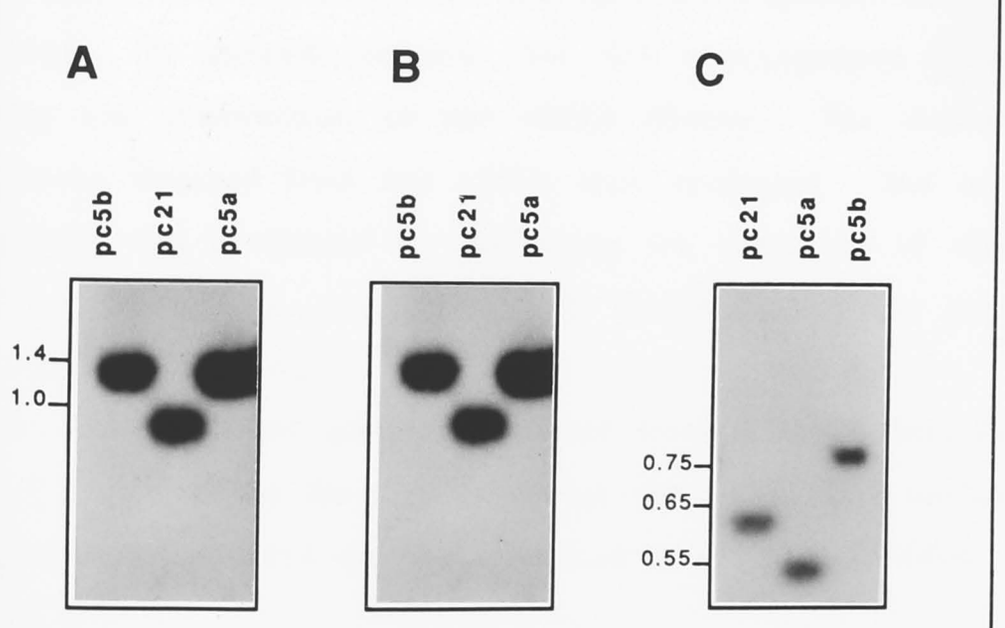


Fig. 4.2 Southern analysis of the *A112* cDNAs.
Lambda DNAs containing the inserts of the *A112* cDNAs were cut with *Eco* RI and hybridized to the following radioactive probes: A the large fragment of pc5a, B the genomic fragment p37-23 and C the small fragment of pc5a. The sizes of the bands in autoradiogram A and B are 1.4kb and 1.0kb, corresponding to the large fragments of the cDNAs. The sizes of the signals in autoradiogram C are 0.55, 0.65 and 0.75kb.

fragment, the small fragments of the cDNAs correspond to the adjacent genomic 4.5kb *Eco* RI fragment; there is no cross hybridization between the large and the small fragments.

4.3.2 The *A112* nucleotide sequence

The sequence information was assembled in three steps. First the sequence of the pc37-3.8 cDNA was obtained using overlapping subclones. By comparison of the cDNA clone pc37-3.8 with the genomic clones it was found that this cDNA contained a re-arrangement. As a result the internal *Eco* RI site appeared at the 3' and 5' ends of this cDNA (*Eco* RI site at nucleotide 1414, Fig. 4.3). This means that effectively, the two *Eco* RI fragments were each inverted. It seemed probable that this rearrangement occurred during the construction of the cDNA library. The nucleotide sequences obtained from this cDNA were re-aligned. The correct alignment was confirmed by comparing the sequences of cDNAs and genomic DNAs, which had been obtained using the primers indicated in figure 4.3.

Note: The cDNA and genomic sequences compared were from the 5' and 3' region of the the cDNA only i.e. the region from nucleotide 732 to nucleotide 1414 was not sequenced in the genomic DNA.

As a second step, another set of cDNAs was isolated and these were compared to the sequenced cDNA pc37-3.8.

Third, missing sequences were obtained using specifically designed primers. This strategy is illustrated in figure 4.3. The random subclones originating from cDNA pc37-3.8, depicted in the top part of the figure, comprise the nucleotide sequence from basepair 500 to 2000. To illustrate the relationships of pc37-3.8 to

the new cDNAs their restriction maps are aligned relative to the genomic *Eco* RI restriction site at position -10 depicted in figure 2.1 (chapter 2). The positions of the various primers used in the sequencing are indicated as black squares.

The assembled nucleotide sequence of the cDNAs corresponding to the *A112* transcript is 1.95kb long (Fig. 4.4). This sequence is about 250bp shorter than the single transcript of 2.2kb that was observed on the northern blots (Fig. 3.1, chapter 3).

A comparison of the restriction maps of the cDNAs and genomic DNA reveals two introns. One intron, of approximately 300bp, is within the genomic *Eco* RI/*Xho* I fragment from position -10 to -9 because this fragment was mapped as 1kb, but the size derived from the cDNA sequence is 682bp (Fig. 4.4). A second intron of 60bp is located 130bp downstream to the 5' end and 38bp downstream to the primer termed PAR1 (Fig. 4.4).

Translation of the nucleotide sequence into the putative amino acid sequence shows only one possible open reading frame, ending with the first stop codon at nucleotide 1750. A poly(A) tail is present 200bp downstream from the first stop codon and 45 bp downstream of the polyadenylation signal AATAAA (underlined in Fig. 4.4). This interpretation is supported by the fact that there is no string of adenosines in the genomic sequence and also by the fact that the number of adenosines varies between different cDNAs.

The usual distance from the poly-(A) signal to the poly-(A) tail is 10 to 30 nucleotides (Proudfoot and Brownlee, 1976; Birnstiel *et al.*, 1985; Proudfoot, 1991). Within this area there are two possible candidates for a poly-(A) signal if a one base pair change is

Figure 4.3

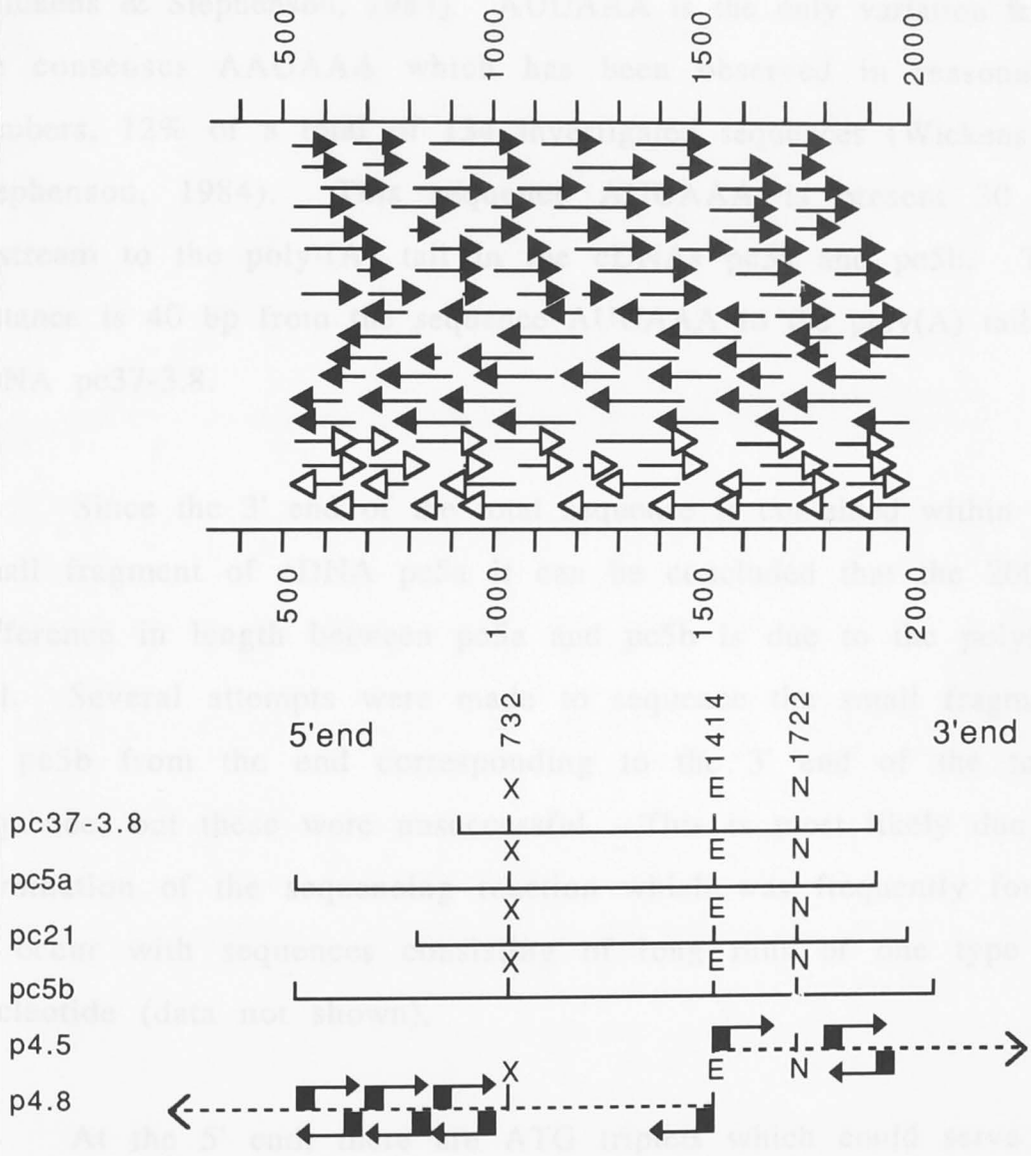


Fig. 4.3 Sequencing strategy. The top part illustrates the overlapping subclones of the cDNA pc37-3.8 which comprise the final *A112* sequence from nucleotide 500 to nucleotide 1950. Each arrow represents a subclone, sequenced several times with GTP (filled arrow heads) and ITP (empty arrow heads). Below are the restriction maps of the cDNAs pc37-3.8, pc5a, pc21 and pc5b, aligned at the internal *Eco* R1 restriction site (E) which corresponds to the genomic *Eco* R1 site at position -10 in the co-ordination system of figure 2.1 (N=*Nhe* I, X=*Xho* I) Indicated above each restriction site is the nucleotide position corresponding to the sequence in figure 4.4. cDNAs and genomic clones were also sequenced with primers, presented as filled boxes, the direction of the sequencing reaction is indicated with arrows. Only parts of the genomic fragments p4.5 (from position -10 to -14.5) and p4.8 (from position -10 to -5.2) are depicted (dashed arrows). Note that in order to present the cDNAs with the 5'end at the left and the 3'end at the right, the orientation of the restriction maps had to be presented inverse to the co-ordination system of the genomic maps in chapter 2.

tolerated (underlined in Fig. 4.4). It has been shown however, that changes in the poly(A) signal drastically affect the mRNA stability (Wickens & Stephenson, 1984). AUUAAA is the only variation from the consensus AAUAAA which has been observed in reasonable numbers, 12% of a total of 134 investigated sequences (Wickens & Stephenson, 1984). This sequence AUUAAA is present 30 bp upstream to the poly-(A) tail in the cDNAs pc5a and pc5b. The distance is 40 bp from the sequence AUUAAA to the poly(A) tail in cDNA pc37-3.8.

Since the 3' end of the total sequence is contained within the small fragment of cDNA pc5a it can be concluded that the 200bp difference in length between pc5a and pc5b is due to the poly(A) tail. Several attempts were made to sequence the small fragment of pc5b from the end corresponding to the 3' end of the total sequence, but these were unsuccessful. This is most likely due to termination of the sequencing reaction which was frequently found to occur with sequences consisting of long runs of one type of nucleotide (data not shown).

At the 5' end, there are ATG triplets which could serve as initiation codons. In frame with the amino acid sequence there is a methionine at position 80 which could be the possible start. Alternatively, the start could be at the upstream methionine at position 72. In one case the upstream sequence agrees with the consensus sequence for eukaryotic initiation sites of translation, with a purine at position -3 (Kozak, 1981) (Fig. 4.4); in the second case the purine is at -2. Because there is no stop codon upstream to either methionine it is theoretically possible that the cDNA is not quite full length and starts with a methionine further upstream to the sequenced region.

Figure 4.4 Nucleotide sequence and predicted amino acid sequence of the 2.2kb transcript

```
   2>>>>>>>>>>>>>>>
   TTT TTT CTT GTT GGG CCA GAA AAC GTG CGC TGC CAT TTT CCA GCG TTT TTT TTT
    F   F   L   V   G   P   E   N   V   R   C   H   F   P   A   F   F   F

  56                                  >>>>>>>>>>>>>>>PARI>>>>>>>>>>>>
   ATA CCG TCG CCC ACC ATG AGC CAG ATG ACC CAG AAA ACG GTG CAA TTC CAC GAG
    I   P   S   P   T   M   S   Q   M   T   Q   K   T   V   Q   F   H   E
                        72              80
 110
   CTG GAG CTG GAT CAG CGC ATC CTT AAG GCG GTG GCC CAG CTG GGT TGG CAG CAG
    L   E   L   D   Q   R   I   L   K   A   V   A   Q   L   G   W   Q   Q

 164                                                  >>>>>>>>>>>PARII>>>>>>>
   CCC ACA CTT ATC CAG TCG ACG GCG ATA CCG CTG CTT CTG GAG GGC AAG GAT GTG
    P   T   L   I   Q   S   T   A   I   P   L   L   L   E   G   K   D   V

 218>>>>>>>
   GTG GTG CGC GCC CGC ACC GGA TCC GGC AAG ACT GCC ACC TAC GCC CTG CCA CTG
    V   V   R   A   R   T   G   S   G   K   T   A   T   Y   A   L   P   L

 272
   ATT CAG AAG ATC CTT AAC TCC AAA CTG AAC GCC AGC GAG CAG TAT GTG AGC GCT
    I   Q   K   I   L   N   S   K   L   N   A   S   E   Q   Y   V   S   A

 326                                                          <<<<<<<<<<
   GTG GTC CTG GCG CCC ACC AAG GAG TTA TGC CGG CAG TCC CGT AAG GTG ATC GAG
    V   V   L   A   P   T   K   E   L   C   R   Q   S   R   K   V   I   E

 380<<PAFII<<<<<<<<<<<<
   CAG CTG GTC GAG TCC TGC GGC AAG GTT GTA CGC GTG GCG GAC ATT GCC GAT AGC
    Q   L   V   E   S   C   G   K   V   V   R   V   A   D   I   A   D   S

 434
   TCC AAC GAC ACG GTG ACC CAG CGG CAC GCC TTG TCT GAA AGT CCC GAC ATT GTG
    S   N   D   T   V   T   Q   R   H   A   L   S   E   S   P   D   I   V

 488
   GTT GCC ACA CCC GCC AAT TTG CTG GCC TAC GCC GAG GCT GGT AGC GTG GTG GAT
    V   A   T   P   A   N   L   L   A   Y   A   E   A   G   S   V   V   D

 542
   CTG AAG CAT GTG GAG ACC TTG GTG GTG GAC GAG GCG GAT CTG GTG TTT GCC TAT
    L   K   H   V   E   T   L   V   V   D   E   A   D   L   V   F   A   Y

 596      <<<<<<<<<<<PAFI<<<<<<<<<<<<<
   GGC TAC GAA AAG GAC TTC AAG CGG TTG ATT AAG CAC CTG CCT CCC ATC TAC CAG
    G   Y   E   K   D   F   K   R   L   I   K   H   L   P   P   I   Y   Q

 650
   GCT GTC CTA GTC TCC GCT ACT CTA ACC GAC GAT GTG GTG CGC ATG AAA GGC CTG
    A   V   L   V   S   A   T   L   T   D   D   V   V   R   M   K   G   L

 704                                        Xho I.
   TGC CTG AAC AAC CCG GTT ACG CTT AAA CTC GAG GAG CCG GAG TTG GTG CCG CAA
    C   L   N   N   P   V   T   L   K   L   E   E   P   E   L   V   P   Q

 758
   GAT CAG TTG TCG CAC CAG CGC ATT CTG GCT GAG GAA AAC GAT AAA CCG GCT ATC
    D   Q   L   S   H   Q   R   I   L   A   E   E   N   D   K   P   A   I

 812
   TTG TAC GCC TTA TTG AAG CTG AGA CTC ATA CGC GGC AAG AGC ATA ATC TTT GTG
    L   Y   A   L   L   K   L   R   L   I   R   G   K   S   I   I   F   V

 866
   AAC AGC ATC GAT CGG TGT TAC AAG GTT CGA TTG TTC CTA GAG CAG TTC GGC ATC
    N   S   I   D   R   C   Y   K   V   R   L   F   L   E   Q   F   G   I
```

Figure 4.4 continued

```
920
   CGT GCC TGT GTC TTG AAC TCC GAG CTG CCT GCC AAC ATT CGC ATC CAC ACA ATT
    R   A   C   V   L   N   S   E   L   P   A   N   I   R   I   H   T   I

974
   AGC CAG TTC AAC AAG GGC ACC TAC GAC ATA ATC ATT GCC TCC GAC GAA CAT CAT
    S   Q   F   N   K   G   T   Y   D   I   I   I   A   S   D   E   H   H

1028
   ATG GAA AAG CCA GGA GGC AAA TCA GCG ACT AAC CGA AAA TCT CCT CGA AGC GGC
    M   E   K   P   G   G   K   S   A   T   N   R   K   S   P   R   S   G

1082
   GAC ATG GAG TCG AGT GCT TCC CGC GGC ATT GAC TTT CAG TGC GTG AAC AAC GTA
    D   M   E   S   S   A   S   R   G   I   D   F   Q   C   V   N   N   V

1136
   ATC AAC TTC GAC TTC CCC AGG GAT GTC ACG TCT TAT ATC CAT CGG GCT GGC AGG
    I   N   F   D   F   P   R   D   V   T   S   Y   I   H   R   A   G   R

1190
   ACG GCT AGA GGA AAT AAC AAG GGC TCC GTC TTG TCC TTT GTC AGC ATG AAG GAG
    T   A   R   G   N   N   K   G   S   V   L   S   F   V   S   M   K   E

1244
   TCC AAG GTA AAC GAT TCA GTT GAG AAG AAA CTG TGC GAT AGT TTT GCA GCC CAA
    S   K   V   N   D   S   V   E   K   K   L   C   D   S   F   A   A   Q

1298
   GAG GGC GAA CAG ATC ATC AAG AAC TAC CAG TTT AAA ATG GAA GAA GTT GAG TCC
    E   G   E   Q   I   I   K   N   Y   Q   F   K   M   E   E   V   E   S

1352
   TTC CGT TAT CGT GCT CAG GAT TGC TGG CGG GCC GCA ACT CGC GTA GCT GTT CAC
    F   R   Y   R   A   Q   D   C   W   R   A   A   T   R   V   A   V   H

1406          Eco RI.
   GAC ACT CGA ATT CGA GAG ATA AAG ATA GAG ATC CTC AAC TGC GAG AAA CTA AAG
    D   T   R   I   R   E   I   K   I   E   I   L   N   C   E   K   L   K

1460
   GCA TTT TTC GAG GAG AAC AAA CGC GAT CTG CAA GCG CTT CGG CAC GAC AAG CCT
    A   F   F   E   E   N   K   R   D   L   Q   A   L   R   H   D   K   P

1514
   CTG CGC GCC ATC AAG GTA CAG AGT CAT CTC TCT GAC ATG CCC GAG TAC ATA GTG
    L   R   A   I   K   V   Q   S   H   L   S   D   M   P   E   Y   I   V

1568
   CCA AAG GCC CTG AAG CGA GTG GTT GGA ACG TCC TCT TCC CCT GTC GGA GCC TCG
    P   K   A   L   K   R   V   V   G   T   S   S   S   P   V   G   A   S

1622
   GAA GCC AAA CAG CCA CGA CAG TCG GCC GCC AAG GCT GCC TTC GAG CGG CAG GTC
    E   A   K   Q   P   R   Q   S   A   A   K   A   A   F   E   R   Q   V

1676                  Nhe I.
   AAT GAT CCT CTG ATG GCT AGC CAG GTG GAC TTC GGG AAA CGG CGT CCT GCC CAC
    N   D   P   L   M   A   S   Q   V   D   F   G   K   R   R   P   A   H

1730                          Stop
   CGG AGA AAA AAG AAG GCG TTG TAG GGT AGC AAA TGG TAT CTA TAG TCG CAT TAG
                              *                   *           *

1784
   TCA ATA AGG AAA TAA AGT TAA TAT TGT TAA TGT AAG CCA AAC ATA AAA GCG GCC
    S           *       *           *
```

Figure 4.4 continued

```
1838
   GTT GAA GAT GTT TGG AAT TTT AAA AAT GTT ATT TTA CCC TAA ATT AGA AAT GAG
    V   E   D   V   W   N   F   K   N   V   I   L   P   *   I   R   N   E

1892   Poly(A)
   TAA AAT TAA ACG AAA AAT ATA CCT ATG TTG GAG ATA GCT GCA GAT TGA CAA ACC
    *       *                                                   *

1946
   CAA AAA AAA AAA AAA
```

Figure 4.4 The nucleotide sequence of the putative *A112* transcript and the translation into the predicted amino acid sequence. The nucleotide positions are indicated to the left of each lane. The possible poly(A) signals (at 1794, at 1897 and at 1908) are underlined. The possible start methionines are highlighted (position 72 and 80). The primers (corresponding to the primers in figure 4.3) used to establish the sequence are indicated as arrows with the tip of the arrows pointing in the direction of the sequencing reaction. The restriction sites *Eco* RI, *Xho* I and *Nhe* I (corresponding to the restriction sites in figure 4.3) are indicated. The translation into the predicted amino acid sequence is presented in one letter code. The only long open reading frame is shown starting at nucleotide 2, ending with the first stop codon at 1750.

Sequencing of several clones derived from the cDNAs pc5a and pc5b revealed a one base pair difference between the two cDNAs: pc5b, like the genomic sequence, contains a cytidine at nucleotide 348 instead of a thymidine which is present in pc5a. The difference has no effect on the putative amino acid sequence.

4.3.3 The predicted A112 protein

The translated putative amino acid sequence results in a protein of 583 amino acids. A search of the Swiss-Prot (release 14.0) and GenPept protein data libraries (release 63.0) for similar sequences using the FASTA algorithm (Pearson & Lipman, 1988) revealed a striking similarity of the predicted A112 amino acid sequence to members of a recently defined RNA helicase family. The fifteen sequences with the highest homology were each specified as putative ATP-dependent RNA helicases. The similarity of the *A112* gene sequence to the members of the RNA helicase family ranges from a 40% identity in a 60 amino acid overlap for the *D. melanogaster* gene *rm62* (Dorer *et al.*, 1990), to 25% identity in an 362 amino acid overlap of the yeast gene *prp5* (Dalbadie-McFarland and Abelson, 1990). These best scores are listed in table 4.1.

The members of this family are very diverse. For instance, two genes *prp5* and *mss116* are involved in splicing processes. PRP5 is a protein directly involved in nuclear splicing and spliceosome assembly (Dalbadie-McFarland and Abelson, 1990), while the yeast protein MSS116 is focussed on the splicing of mitochondrial transcripts (Seraphin *et al.*, 1989). The P68 protein on the other hand is a major nuclear antigen from a dividing

human hepatoma which is thought to be important in the regulation of cell growth and division (Ford *et al.*, 1988). Some

Table 4.1

protein	aa position	% ID	N	references
RM62	340 - 400	40	459	Dorer *et al.*, 1990
SRMB	20 - 340	31	410	Nishi *et al.*, 1988
PRP5	20 - 380	25	346	Dalbadie-McFarland & Abelson, 1990
P68	340 - 390	45	339	Ford *et al.*,1988
TIF	155 - 395	21	326	Linder & Slonimski, 1989
PL10	300 - 400	26.5	315	Leroy *et al.*, 1989
DBPA	270 - 400	22	312	Iggo *et al.*, 1990
eIF-4A-II	310 - 400	31	302	Nielsen & Trachsel, 1988
eIF-4A-I	310 - 400	29	298	Nielsen & Trachsel, 1988
vasa	20 - 310	23	274	Lasko & Ashburner, 1988 Hay *et al.*, 1988
MSS116	20 - 400	25	255	Seraphin *et al.*, 1989
HCA4	160 - 310	26	163	Chang *et al.*, 1990
HCA3	340 - 370	53	146	Chang *et al.*, 1990

Table 4.1 List of amino acid sequences with highest scores retrieved from the Swissprot and Genpept databank. The similarities of the predicted A112 protein to the best scoring proteins (column 1) are listed in percent identity (% ID in column 3) within an amino acid stretch specified by its amino acid position (aa position in column 2) ranked by the optimal score (N) in column 4.

members of the RNA helicase family are involved in translation, for example TIF1 and TIF2 are thought to be the translation initiation factors in yeast (Linder and Slonimski, 1989) analogous to eIF-4A, the translation initiation factor in the mouse (Nielsen *et al.*, 1985;

Nielsen and Trachsel, 1988). PL10 is a protein expressed specifically during spermatogenesis in mouse (Leroy *et al.*, 1989) and the gene product of *vasa* is involved in oogenesis and specification of the anterior-posterior axis of *D. melanogaster* embryos (Lasko and Ashburner, 1988). Two RNA helicases have been discovered so far in *E.coli*, one is *dbpA* (Iggo *et al.*, 1990) the other is *srmB* (Nishi *et al.*, 1988). If *srmB* is present in high gene dosage, it suppresses a mutation in the L24 ribosomal protein, which is essential for the assembly of 50S ribosomal subunits in *E. coli* (Nishi *et al.*, 1988). The protein RM62 was discovered in an attempt to find the genes corresponding to the *Triplo-lethal* locus in *D. melanogaster* (Dorer *et al.*, 1990).

Although the biological functions are very different for each of those proteins, they share a sequence of approximately 400 amino acids with significant homology. In some proteins like A112 and SrmB the 400 residues constitute the main part of the protein, others like PRP5 and the *vasa* protein, extend considerably beyond these 400 amino acids (see N-terminal and C-terminal sequences in Table 4.2). Table 4.2 shows the similarity of the *A112* gene to the members of the RNA helicase family based on a comparison of these 400 residues rather than the alignment of the entire amino acid sequences. This comparison reveals that the sequence similarity is centred around two domains with high homology. One is at the beginning of the *A112* sequence, from amino acid 30 to 200 (Table 4.2a). This domain contains the first four of the seven motifs noted by Gorbalenya (1988) and Hodgman (1988) which define the helicase family (Table 4.2, also underlined in Fig. 4.4).

The second domain is at the end of the predicted A112 protein, from amino acid number 340 to 380 (Table 4.2) and

contains the motifs V and VI (Gorbalenya *et al.*, 1988; Hodgman 1988).

Parts of these domains are assigned to possible functions. For example, the sequence D X_4 A X_4 GKT (Table 4.2, motif I) is typical for the A-motif of ATP binding proteins (Walker *et al.*, 1982; Gorbalenya *et al.*, 1988; Hodgman 1988; Linder *et al.*, 1989). The analysis of the predicted A112 amino acid sequence demonstrates that conserved amino acids of the A-motif and surrounding sequences are present in the A112 protein (Table 4.2). Additionally the sequence (V/I) L D E A D (M/L) L X_2 G F, or DEAD- box is present (Table 4.2, motif II). This DEAD- box has been established as a special version of the B-motif of ATP-binding proteins (Walker *et al.*, 1982; Linder *et al.*, 1989).

Both motifs are thought to play a role in unwinding RNA/DNA helices under ATP hydrolysis. The A-motif is thought to bind one of the phosphates of ATP. The second phosphate is thought to be bound by an aspartate via a magnesium ion which is part of the B-motif of ATP-binding proteins (Walker, 1982; Hodgman, 1988). The DEAD- box is an analogous version of the B-motif which is found uniquely in RNA helicases (Linder *et al.*, 1989; Chang *et al.*, 1990). The equivalent sequence to DEAD in DNA helicases is $^{R}/_{K}$-X_3-G-X_3-L -(hydrophobic)$_4$-aspartate, where the four hydrophobic residues plus the aspartate represent the equivalent environment to the sequences around the DEAD box (Walker *et al.*, 1982). In the predicted A112 protein both motifs are connected with a sequence of considerable similarity and comparable spacing to the sequence of known RNA helicases.

The second domain from amino acid 340 to 380 contains the HRIGR motif which is equally conserved in RNA helicases (Nielsen and Trachsel, 1988; Chang *et al.*, 1990; Seraphin *et al.*, 1990). Although a particular function of this motif has not been verified, there is speculation that it is involved in polynucleotide binding (Linder *et al.*, 1989). The A112 amino acid sequence is virtually identical to the consensus sequence established from members of the helicase gene family, the only noteworthy difference is an alanine (A) at amino acid position 394 rather than an isoleucine (I) (Table 4.2, motif VI; Fig. 4.4).

The spacing between domain one and domain two (Table 4.2) varies from 114 residues in *vasa* to 136 amino acids in the PRP5 protein. In the DNA/RNA helicase family originally defined by Gorbalenya (1988) and Hodgman (1988) this spacing contains an additional motif denoted motif IV. Within those two domains the spacing between the motifs remains constant throughout the sequence (Table 4.2).

Table 4.2

	N						D	-X_4-	A	-X_4-	GKT		$APTX_2L$
							motif I					motif Ia	
A112	29	**IQ**	STAI	P	4	GK	**D**	VVVR	**A**	RTGS	**GKT**	28	APTKEL
RM62	162	**IQ**	AQGW	P	4	GS	N	FVGI	**A**	KTGS	**GKT**	28	APTREL
SRMB	29	**IQ**	AAAI	P	4	GR	**D**	VLGS	**A**	PTGT	**GKT**	27	TPSSRA
PRP5	281	**IQ**	SQAL	P	4	GR	**D**	VIGI	S	KTGS	**GKT**	29	APTREL
P68	99	**IQ**	AQGW	P	4	GL	**D**	MVGV	**A**	QTGS	**GKT**	28	APTREL
TIF	47	**IQ**	QRAI	M	4	GH	**D**	VLAQ	**A**	QSGT	**GKT**	23	APTREL
PL10	204	**VQ**	KHAI	P	4	KR	**D**	LMAC	**A**	QTGS	**GKT**	41	APTREL
DBPA	3	**VQ**	AAAL	P	4	GK	**D**	VRVQ	**A**	KTGS	**GKT**	23	CPTREL
eIF-4A	58	**IQ**	QRAI	I	4	GY	**D**	VLAQ	**A**	QSGT	**GKT**	23	APTREL
vasa	270	**IQ**	KCSI	P	4	GR	**D**	LMAC	**A**	QTGS	**GKT**	28	SPTREL
MSS116	131	**VQ**	QKTI	K	6	DH	**D**	VIAR	**A**	KTGT	**GKT**	27	APTRDL

				$^V/_IL$	DEAD	X_N			Q	X	VL	X	SAT
		motif II							motif III				
A112	43	**IVVATPA**	20	**V V**	**DEAD**	5	GY	15	Q	A	VL	V	**SAT**
RM62	40	**IVIATPG**	19	**VL**	**DEAD**	5	**GF**	15	Q	T	LM	W	**SAT**
SRMB	41	**IVVATTG**	19	**I L**	**DEAD**	5	**GF**	15	Q	T	LL	F	**SAT**
PRP5	41	**IVVATPG**	22	**V** M	**DEAD**	5	**GF**	15	Q	C	VL	F	**SAT**
P68	40	**ICIATPG**	19	**V L**	**DEAD**	5	**GF**	15	Q	T	LM	W	**SAT**
TIF	39	**IVVGTPG**	19	**I L**	**DEAD**	5	**GF**	15	Q	V	VL	L	**SAT**
PL10	40	LL**VATPG**	19	**V L**	**DEAD**	5	**GF**	19	H	T	MM	F	**SAT**
DBPA	41	II**VATPG**	19	**V** M	**DEAD**	5	**GF**	15	Q	T	LL	F	**SAT**
eIF-4A	41	**IVVGTPG**	19	**V L**	**DEAD**	5	**GF**	15	Q	V	VL	L	**SAT**
vasa	40	**VVIATPG**	19	**V L**	**DEAD**	5	**GF**	17	Q	T	LM	F	**SAT**
MSS116	45	**IVIATPG**	20	**V L**	**DEAD**	5	**GF**	22	K	T	LL	F	**SAT**

Table 4.2 Alignment of conserved motifs from 10 amino acid sequences to demonstrate their similarity with the A112 protein. Positions of similar or identical amino acids are presented. The number of amino acids between two blocks of high similarity is shown. N indicates the amino acids at the N-terminal end of the proteins, C indicates the C-terminal end. Amino acids are highlighted when they are identical with the motif indicated at the top of each column.

		motif V				motif VI	
protein	X_N	ASRGID	X_N	VIN$^{F}/_{Y}$D	X_N	YIHR I GRT X R	-C
A112	142	ASRGID	6	VIN F D	7	YIHR AGRT AR	183
RM62	121	AARGLD	6	VIN F D	7	YIHR IGRT GR	99
SRMB	117	AARGID	6	VFN F D	7	YLHR IGRT AR	181
PRP5	136	L SRGLN	6	VII Y N	7	YVHT TGRT AR	231
P68	119	ASRGLD	6	VIN Y D	7	YIHR IGRT AR	179
TIF	116	LARGID	6	VIN Y D	7	YIHR IGRG GR	42
PL10	116	AARGLD	6	VIN F D	7	YVHR IGRT GR	127
DBPA	114	AARGLD	6	VVN FE	7	HVHR IGRT AR	123
eIF-4A	116	LARGID	6	VIN Y D	7	YIHR IGRG GR	41
vasa	114	ASRGLD	6	VIN Y D	7	YVHR IGRT GC	80
MSS116	128	GARGMD	6	VLQ I G	7	YIHR IGRT AR	186

Table 4.2 continued Alignment of conserved motifs from 10 amino acid sequences to demonstrate their similarity with the A112 protein. Positions of similar or identical amino acids are presented. The number of amino acids between two blocks of high similarity is shown. N indicates the amino acids at the N-terminal end of the proteins, C indicates the C-terminal end. Amino acids are highlighted when they are identical with the motif indicated at the top of each column.

4.4 Discussion

The *A112* nucleotide sequence

A nucleotide sequence was obtained from cDNA clones corresponding to a genomic area to which the *A 1 1 2* complementation group had been related (see chapter 3). The 3' end of the gene was identified by the poly-A tail in the cDNA sequences and by the presence of a poly adenylation signal. The presumed 5' end was determined by the onset of a continous reading frame starting with one of two possible methionines at positions 72 and 80 respectively. The putative *A112* sequence contains at least two introns. Since genomic sequences were established using primers, the genomic sequence is not contiguous and therefore it cannot be excluded that there are other introns besides the two discovered. This uncertainty could be removed by sequencing the corresponding genomic fragment. This might also determine whether or not the cDNA is full length. It is possible that the cDNA might not be quite full length because there is no stop codon prior to the first methionine (Fig. 4.4). A comparison of the cDNA length with the size of the *A112* transcript on Northern blots (2.2kb in Fig. 3.14, chapter3) shows that up to 250 nucleotides could be missing. However this difference in length could be due to the string of adenosines in the poly-(A) tail at the 3' end.

The upstream region could be defined by an S1 nuclease protection assay using a genomic fragment that contains the 5' end of the cDNA sequence, for example the 4.8kb *Eco*R1 fragment from position -10 to -5.2. In such an experiment whole *D. melanogaster*

RNA is hybridized with the radiolabeled DNA fragment and subsequently the single stranded RNA or DNA is digested by nuclease S1, that does not affect double stranded nucleotides. The size of the hybrid fragment is analysed to see if a larger RNA sequence is protected than the fragment of the known cDNA sequence included in the genomic probe. This method will reveal the presence of even small exons upstream of the sequenced cDNA.

Another sensitive method, RACE (rapid amplification of cDNA ends), uses primer extension combined with PCR (polymerase chain reaction) (Frohman *et al.*, 1988). An antisense oligonucleotide from an exon at the 5' end of the known cDNA is used as a primer for reverse transcriptase. Whole RNA is utilized as a template and the first strand product is tailed with a homopolymer by terminal transferase. In the next step, a hybrid primer, consisting of the complementary homopolymer sequence and a random sequence, is used to create a second strand with reverse transferase utilizing the first strand cDNA as a template. By using the PCR reaction the second strand product can be amplified up to 10^6 copies of cDNA using the random sequence of the hybrid primer and the oligo nucleotide as primers (Sambrook *et al.*, 1989). The size of the amplified cDNA is checked by gel electrophoresis and southern transfer after the excess single strand DNA has been removed by digestion with mung bean nuclease. The size of the amplified cDNA is then compared to the known cDNA sequence and by difference in length, even a small exon should be detected.

Using these methods together with genomic sequencing upstream of the currently available clones it should be possible to produce a map of the intron/exon structure of the *A112* gene. Within the time constraints of the PhD course I was unable to persue any of these suggested experiments.

The predicted A112 protein

The predicted A112 amino acid sequence has features in common with members of a recently defined gene family, the RNA helicase gene family. This gene family, consisting of both established and putative RNA helicases, was defined on the basis of the amino acid sequence alignments of their proteins (Gorbalenya, *et al.*, 1988; Hodgman, 1988; Linder *et al.*, 1989). Any alignment of two proteins will show some accidental homology which complicates the interpretation of the significance of weakly homologous sequences. However, alignment of several proteins with each other avoids this problem, because it is much less likely that several proteins will have the same conserved sequence.

The motifs noted by Hodgman are a feature of a very broad group of helicases (Lane, 1988). Since the discovery of the original motifs an increasing number of helicases have been incorporated into the comparison leading to finer alignment of conserved residues within the helicase family and further division into subfamilies based on the differences in the conserved motifs (Gorbalenya *et al.*, 1988; Hodgman, 1988; Gorbalenya *et al.*, 1989). One of these subfamilies is the RNA helicase family which is characterized by the DEAD box (Linder *et al.*, 1989). This motif changes to DEAH in another subfamily recently characterized which is composed of proteins involved in pre mRNA splicing (Burgess *et al.*, 1990; Chen and Lin, 1990; King and Beggs, 1990; Company *et al.*, 1991; Kuroda *et al.*, 1991; Schwer and Guthrie, 1991).

The members of the RNA helicase family include proteins from *E.coli* (SrmB and DbpA), yeast (TIF1, TIF2, MSS116), *D. melanogaster* (*vasa*), mouse (*eIF-4A*, *PL10*) and human (*eIF-4A*, *p68*). Overall, the amino acid sequences share about 30% homology

(Ford *et al.*, 1988; Lane, 1988; Nishi *et al.*, 1988; Linder and Slonimski, 1989; Seraphin *et al.*, 1989; Iggo *et al.*, 1990). The predicted A112 amino acid sequence clearly belongs to the family characterized by the DEAD box since it contains the Hodgman/Gorbalenya motifs (Gorbalenya *et al.*, 1988; Hodgman, 1988) including the DEAD box which gives the family its name (Linder *et al.*, 1989). Its overall similarity of 31% to *srmB* and 40% to *rm62* is also in good agreement with the overall similarity of 30% amongst the other members.

The biological function of the putative A112 protein.

It can be postulated that the biological function of the *A112* complementation group is essential for viability since every mutant allele discovered at this locus is homozygous lethal and germline clone analysis revealed that mutations at this locus are germ cell lethal (Perrimon *et al.*, 1989). This is not necessarily a contradiction to the proposed function of a RNA helicase. Although there are numerous RNA helicases to be expected in *D. melanogaster*, each one can still be connected to a specific function which might or might not be connected to viability. In the case of the yeast TIF1 and TIF2 genes, also members of the RNA helicase family, inactivation of either gene by gene disruption has no effect on cell viability or mitochondrial functions. However, simultaneous inactivation of both genes is lethal to the cell (Linder *et al.*, 1989; Chang *et al.*, 1990) suggesting that their function cannot be substituted for by other RNA helicases although 10 have been discovered so far in yeast (Chang *et al.*, 1990).

Within the RNA helicase family, the *A112* gene shows similarities to mRNA splicing genes (the highest degree of similarity was found to PRP5, a protein required for mRNA splicing

in yeast (Dalbadie-McFarland and Abelson, 1990) and *mss116*, a nuclear gene which also affects the splicing of primary transcripts. However, *A112* is most likely not the *D. melanogaster* equivalent to these genes since the overall similarity of *A112* to either amino acid sequence is 25% (Table 4.1). In comparison the predicted amino acid sequence of the *maleless* gene when compared to these proteins has an overall similarity of 36% and is similar to 47% within the domains (Kuroda *et al.*, 1991).

The putative *A112* sequence also shows homology to p68, a protein involved in cell division. However, similar arguments apply as the similarities between p68 and the putative A112 protein is not significantly greater than similarities between p68 and other members of the RNA helicase family. It is therefore not possible to elucidate the function of *A112* further without experiments. The most obvious experiment would be to purify the A112 protein with the intention to test whether it has RNA dependent ATPase activity as was shown for p68 (Hirling *et al.*, 1989), SrmB (Nishi *et al.*, 1988) and Prp 16 (Schwer and Guthrie, 1991).

Chapter 5

Sequencing of the "collagen-like" gene

5.1 Introduction

The data described in chapter 2 suggest that the most likely borders of the *LB20* locus are between -10 and 6.5 (Fig. 2.13) on the genomic map. Three RNA transcripts were detected in total RNA within this area. Firstly a 2.2kb transcript corresponding to the DNA in the area from -13 to -6.8, secondly a 0.6kb transcript corresponding to the DNA position 0 to 1.3 and thirdly a 0.4kb transcript corresponding to the DNA area from 2 to 6.5. One of those, the 2.2kb transcript, can be ruled out as a possible candidate for the *LB20* transcription unit, since it spans the breakpoint of *Df(1)HM44* at position -10, which separates the loci *A112* and *LB20*. The remaining two transcripts are both reasonable candidates to be part of the *LB20* transcription unit, since they exist within the mapped area. However, one would expect to find the *LB20* transcript in close proximity to the proximal breakpoint of *Df(1)JA117*, which had been mapped close to position -5.2. Therefore the 0.6kb transcript appeared to be a more likely candidate than the 0.4kb transcript.

It was important to sequence the transcript corresponding to the genomic fragment from 0 to 1.3, discussed in chapter 3 (Fig. 3.1) as it was closer to the proximal breakpoints of *Df(1)HM44* and *Df(1)JA117* than was the 0.4kb transcript.

Also, this transcript was discovered using a probe which originally had been isolated on the basis of its similarity to a chicken collagen gene probe. The chicken collagen gene probe

hybridized to two *D. melanogaster* clones. One, termed *DCg1*, which was localised to the 25C locus on chromosome 2L, and the other, termed *DCg2*, to the base of the X-chromosome over the region 19EF/20AB (Natzle *et al.*, 1982). The clone *DCg2* was found to hybridize to the 1.3kb *Hind* III fragment at position 0 to 1.3 (Miklos, unpublished data), so this was subcloned (termed *pCog*). Hence the transcript corresponding to this clone *pCog* has the potential of being related to collagens, an important class of molecules found in all metazoan phyla (Adams 1978; reviewed in Bornstein and Sage, 1980).

Because of its unique structural features, collagen has been a focus of structure-function studies. As a component of connective tissue, collagen is also studied for its role in tissue development and differentiation, cell adhesion, cell migration and cell microenvironment (Hay, 1984; Politz and Edgar, 1984; Mirre *et al.*, 1988). The characteristic feature common to all collagen molecules is their triple-stranded helical structure. Three collagen α-polypeptide chains are wound around each other generating a rod-like collagen molecule. On the amino acid level this helical domain is based on the amino acid repeat glycine-X-Y, where X and Y are any non-equivalent amino acids and X stands to a large extent for proline (Ramachandran, 1967; Fessler and Fessler, 1978). This structural feature is characteristic of all collagens. All other features once considered to be characteristic have been found to have exceptions (Adams 1978). Sequencing of the transcript corresponding to *pCog* therefore provides a direct approach to determine whether the protein of the "collagen-like" gene does belong to the collagen family.

5.2 Materials and Methods

The techniques used to accomplish the results described in this chapter have been described already in section 4.2, the materials and methods of chapter 4. The library screened for cDNA clones was made from adult RNA (Poole *et al.*, 1985) .

5.3 Results

5.3.1 The genomic sequence of the "collagen-like" gene

In order to obtain more information about the "collagen-like" gene, its sequence was assembled from genomic DNA and cDNAs. The genomic sequences were acquired applying the shotgun technique (Bankier and Barrell, 1984) i.e., overlapping random fragments of the genomic *Hind* III fragment

Figure 5.1 Sequencing strategy

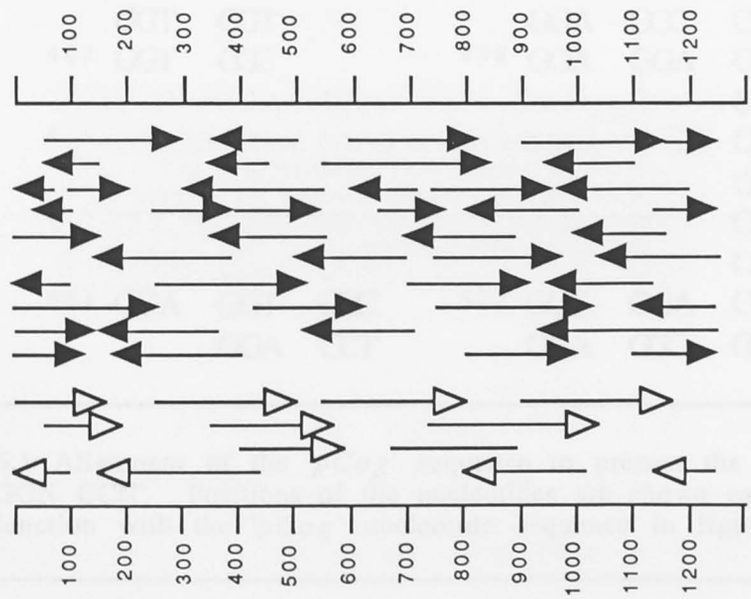
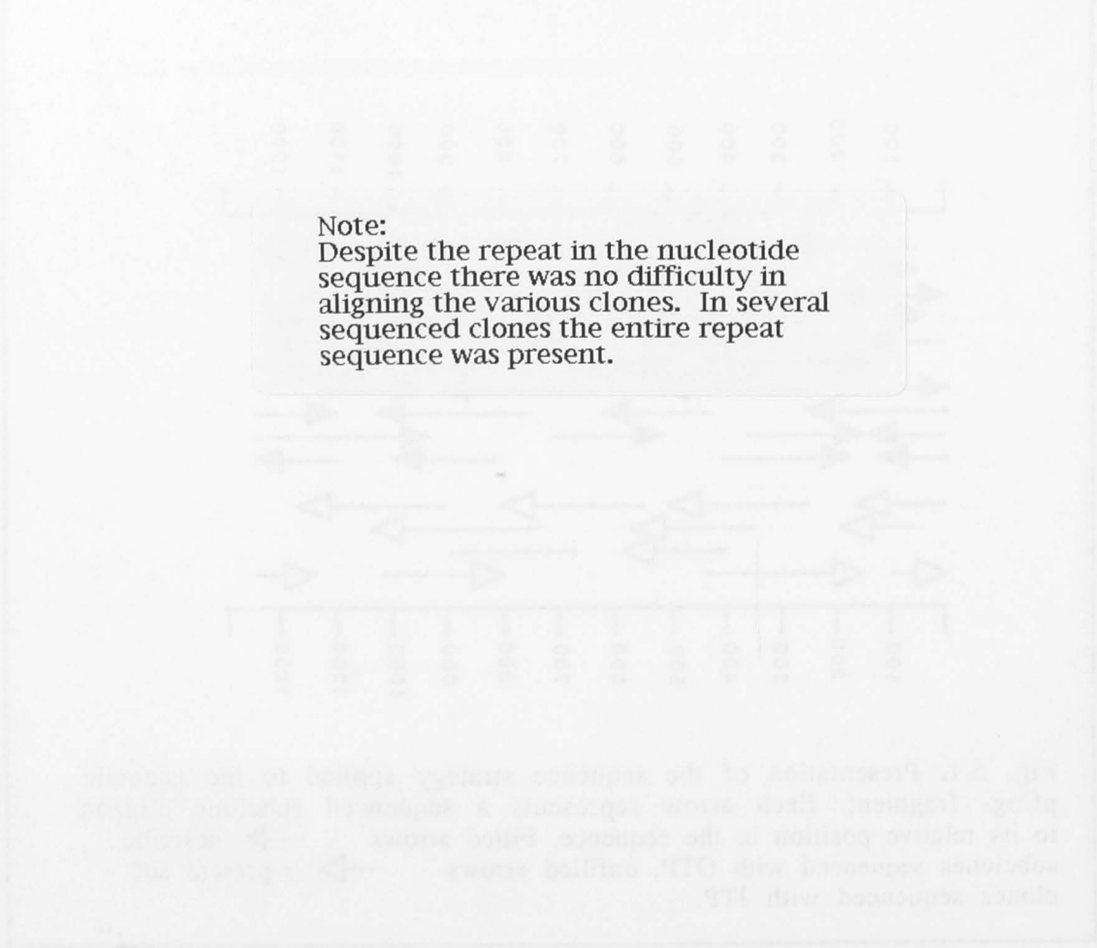


Fig. 5.1. Presentation of the sequence strategy applied to the genomic pCog- fragment. Each arrow represents a sequenced subclone aligned to its relative position in the sequence. Filled arrows → describe subclones sequenced with GTP, unfilled arrows ⇢ represent sub - clones sequenced with ITP.



Note:
Despite the repeat in the nucleotide sequence there was no difficulty in aligning the various clones. In several sequenced clones the entire repeat sequence was present.

(position 0 to 1.3) were created by sonication, ligated into M13-bacteriophage and sequenced as described in chapter 4.2. The sequencing strategy is illustrated in figure 5.1.

Examination of the genomic sequence (Tables 5.1, 5.2, and Fig. 5.4) reveals several interesting features. The most obvious feature is a repetitive sequence. From nucleotide 369 to 568 the sequence contains a repeat consisting almost entirely of the nucleotides guanosine and cytidine (Table 5.1). The basic structure of this repeat is the sequence: 'GGN GGN CCN' with an adenosine as N in 62% of the cases.

Table 5.1

	GGN	GGN	CCN		GGN	GGN	CCN
369	GGT	GGA	CCT	462	GGA	GGA	CCT
	GGT	GGA	CCT		GGA	GGA	CCA
387	GGT	GGA	CCA	480	GGA	GGC	CCA
	GGT	CGT			GGA	GGC	CCA
402	GGT	CGC		498	GGA	GGA	CCA
	GGA	GGA	CCA		GGA	GGA	CCA
417	GGA	CGT		516	GGA	GGC	CCA
		GGA	CCT		GGA	GGA	CCA
429	GGC	GGA	CCA	534	GGA	TGC	CCA
	GGT	GGC	CCA		GGA	GGA	CCA
447	GGA	GGT	CGC	522	GGT	GGA	CCA
		GGA	CCT		GGA	GGC	CCC

Table 5.1 Alignment of the '*pCog*' sequence to present the first repeat: 'GGN GGN CCN'. Positions of the nucleotides are shown as superscripts in conjunction with the '*pCog*' nucleotide sequence in figure 5.4

Almost immediately after this repeat follows a second, less obviously repetitive motif. From nucleotide 643, but more clearly visible from nucleotide 672, the sequence is based on the motif:

'ACN GAN TCN TCN' (Tab 5.2).

Short repetitive sequences, usually a few hundered base pairs in length, have been reported as features of the coding region in a number of genes. For example, in *Drosophila* there is the *opa* repeat (Kidd *et al.*, 1983; McGinnis *et al.*, 1984a; McGinnis *et al.*, 1984b; Wharton *et al.*, 1985) and the *pen* repeat (Digan *et al.*, 1986; Haynes *et al.*, 1987), both of which are simple sequence repeats. The *opa* repeat consists largely of the triplets CAG and CAA, which encode glutamine. The *pen* repeat consists of a variable number of GGN triplets, where N can be any nucleotide.

Table 5.2

	ACN	GAN	TCN	TCN
672	ACA	GAG	TCC	TCT
	ACG	GAA	TCC	TCG
696	ACG	GAA	TCA	TCC
	ACA	GCA	TCC	TCC
720	ACA	GAA		

Table 5.2 Alignment of the '*pCog*' sequence to present the second repeat: 'ACN GAN TCN TCN'. The nucleotide position is given as superscript, in conjunction with the sequence in figure 5.4.

To analyse whether the repeats found in the *pCog* sequence (Tables 5.1 and 5.2) are part of a coding region, the DNA sequence was translated into the amino acid sequence to search for possible open reading frames (Fig. 5.2).

The preliminary translation of the nucleotide sequence into possible amino acid sequences reveals that both repeats are part of each of the three open reading frames which were found (Fig. 5.2).

All three open reading frames are of approximately equal length. ORFI and ORFII are in the same direction, ORFIII is in the opposite orientation.

Figure 5.2

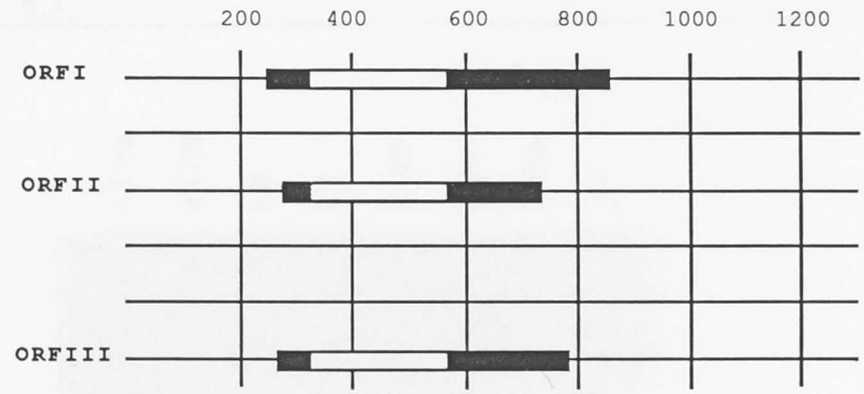


Figure 5.2 Possible open reading frames (designated ORFs I to III) from the sequence of the genomic *pCog* fragment. The six frames possible, when both strands are translated, are symbolised as lines. The numbers at the top represent positions in the nucleotide sequence. The approximate position of the three ORFs greater than 200 amino acids are indicated by solid bars. The relative position of the repeat GGN GGN CCN is presented as an unshaded area within the ORF.

5.3.2 Isolation of cDNAs corresponding to the "collagen-like" gene

To verify the coding region within *pCog*, a cDNA library was screened using the genomic 1.3kb *Hind* III fragment as a probe. A cDNA library made from adult flies (Poole *et al.* 1985) was selected for screening since it was known from the Northern analyses (chapter 3, Fig. 3.1) that the '*pCog*' transcript was very abundant in adult tissue. Eight cDNAs were isolated.

The presence of the repetitive structure within the genomic sequence suggested the possibility that cDNAs unrelated to the gene of interest might be amongst the isolated clones. To test this possibility, the 105bp insert of a M13 bacteriophage clone containing a known non-repetitive part of the genomic sequence, was chosen as a probe (Fig. 5.3). This M13 bacteriophage clone

Figure 5.3

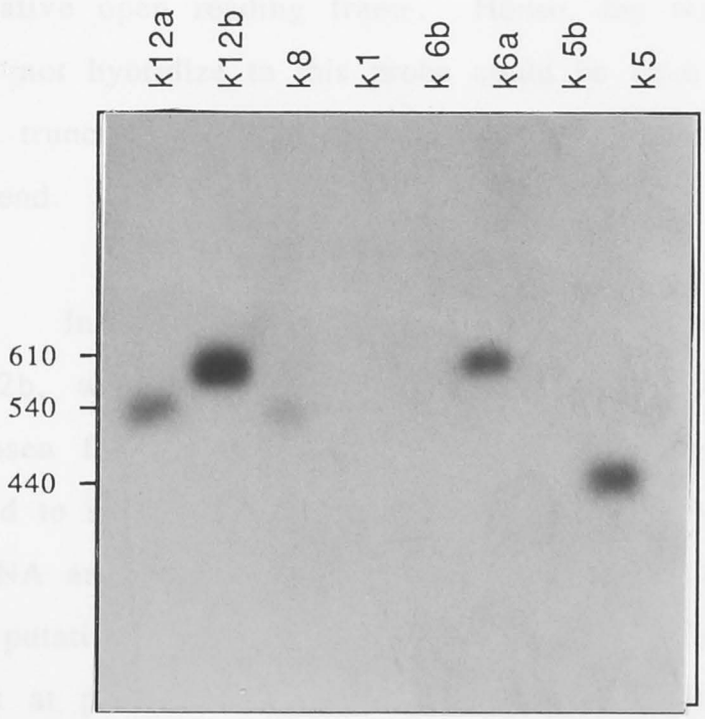


Fig. 5.3 Representative autoradiogram of the southern transfers of eight cDNA inserts, as indicated above each lane. The cDNA inserts were hybridized to a radioactively labelled probe made from the insert of a previously sequenced clone. The sequence of the probe is 100bp long, from nucleotide position 236 to 336. The relative position of the probe in the cDNA and genomic sequence is as indicated in figure 5.4.

Hae114 was one of the sub-clones from the 1.3 *Hind* III fragment that had been sequenced and was known to be upstream of the (GGX) repeat. The 105bp insert of this clone *Hae114*, was hybridized to filters containing the inserts of the isolated cDNA clones (Fig. 5.3). The four cDNAs which hybridized strongly to this 105bp probe (*Hae114*) were chosen for sequence analysis.

The probe (*Hae114*) extends from nucleotide 231 to nucleotide 336, which is about 25 nucleotides upstream from the $(GGX)_n$ repeat (Fig. 5.4). It contains only a very short part of the putative open reading frame. Hence, the remaining cDNAs which did not hybridize to this probe could be from this genomic area but are truncated such that they fail to extend far enough towards the 5' end. They were not examined further.

Initially each clone, subcloned from cDNA K5a, K6a, K12a and K12b, was sequenced from both ends. Additionally a primer chosen from the genomic sequence, (underlined in Fig. 5.4), was used to sequence the total length of the cDNAs. The comparison of cDNA and genomic sequence is presented in figure 5.4. A number of putative transcription features were identified such as the TATA box at position 228, the start signal at position 258 and the poly adenylation signal at position 818 (Fig. 5.4).

Two cDNAs were found to contain a poly(A) tail, thus orienting the direction of transcription. This orientation was confirmed by the presence of a poly(G) run in cDNA K5, which would have been attached to the 5' end during the cDNA construction (Poole *et al.*, 1985). The orientation of transcription agrees with two of the possible open reading frames of the genomic sequence and

Figure 5.4 Alignment of cDNAs and genomic DNA sequences corresponding to the *pCog* fragment

```
pCog    1    AGCTT TGTAA CCAAC AATTT TGGAA ATAAA CCTAA ACATG AATCA ACCAT
pCog    51   TGTCT ACAAA CTCAT ATTCT TTCCA TCTGA CAGAA CCCCA GAGTG AAGGC
pCog    101  CAATG TCAAT TGCGA AGTGA CTGAC TTGGA AAGTG GAATT TACAA ATCCA
pCog    151  CACTG ACGTT AGTTT GCTCA ACGCG CAAAC GTCAC CGATG GAACA CATGT

pCog    201  GCCAC CGAGA CTTTT AGCAG ACGGT GGGTA TATATACCGA TGCCA AGTAA

pCog    251  TGCCG AGATCAGTTC GAGTC GTTAA CTATC AGATA ACGAA ATGCG TACCC

K12b    296                                                       ACCC
K5a     260  GGGGG GGGGG AGTTC GAGTC GTTAA CTATC AGATA ACGAA ATGCG TACCC
K6a     261            GTTC GAGTC GTTAA CTATC AGATA ACGAA ATGCG TACCC
K12a    288                                                 AA ATGCG TACCC

pCog    301  TTATC CTTGT TACTC TTGTT GCCCT GGTCG CAGTG GCCTC CGCCC AAGGA

K12b         TTATC CTTGT TACTC TTGTT GCCCT GGTCG CAGTG GCCTC CGCCC AAGGA
K5a          TTATC CTTGT TACTC TTGTT GCCCT GGTCG CAGTG GCCTC CGCCC AAGGA
K6a          TTATC CTTGT TACTC TTGTT GCCCT GGTCG CAGTG GCCTC CGCCC AAGGA
K12a         TTATC CTTGT TACTC TTGTT GCCCT GGTCG CAGTG GCCTC CGCCC AAGGA

pCog    351  CCTGG TCCTT GGGGT CCCGG TGGAC CTGGT GGACC TGGTG GACCA GGTCG

K12b         CCTGG TCCTT GGGGT CCCGG TGGAC CTGGT GGACC TGGTG GACCA GGTCG
K5a          CCTGG TCCTT GGGGT CCCGG TGGAC CTGGT GGACC TGGTG GACCA GGTCG
K6a          CCTGG TCCTT GGGGT CCCGG TGGAC CTGGT GGACC TGGTG GACCA GGTCG
K12a         CCTGG TCCTT GGGGT CCCGG TGGAC CTGGT GGACC TGGTG GACCA GGTCG

pCog    401  TGGTC GCGGA GGACC AGGAC GTGGA CCTGG CGGAC CAGGT GGCCC AGGAG

K12b         TGGTC GCGGA GGACC AGGAC GTGGA CCTGG CGGAC CAGGT GGCCC AGGAG
K5a          TGGTC GCGGA GGACC AGGAC GTGGA CCTGG CGGAC CAGGT GGCCC AGGAG
K6a          TGGTC GCGGA GGACC AGGAC GTGGA CCTGG CGGAC CAGGT GGCCC AGGAG
K12a         TGGTC GCGGA GGACC AGGAC GTGGA CCTGG CGGAC CAGGT GGCCC AGGAG

pCog    451  GTCGC GGACC TGGAG GACCT GGAGG ACCAG GAGGC CCAGG AGGCC CAGGA

K12b         GTCGC GGACC TGGAG GACCT GGAGG ACCAG GAGGC CCAGG AGGCC CAGGA
K5a          GTCGC GGACC TGGAG GACCT GGAGG ACCAG GAGGC CCAGG AGGCC CAGGA
K6a          GTCGC GGACC TGGAG GACCT GGAGG ACCAG GAGGC CCAGG AGGCC CAGGA
K12b         GTCGC GGACC TGGAG GACCT GGAGG ACCAG GAGGC CCAGG AGGCC CAGGA

pCog    501  GGACC AGGAG GACCA GGAGG CCCAG GAGGA CCAGG ATGCC CAGGA GGACC

K12b         GGACC AGGAG GACCA GGAGG CCCAG GAGGA CCAGG ATGCC CAGGA GGACC
K5a          GGACC AGGAG GACCA GGAGG CCCAG GAGGA CCAGG ATGCC CAGGA GGACC
K6a          GGACC AGGAG GACCA GGAGG CCCAG GAGGA CCAGG ATGCC CAGGA GGACC
K12a         GGACC AGGAG GACCA GGAGG CCCAG GAGGA CCAGG ATGCC CAGGA GGACC
```

```
pCog    551  AGGTG GACCA GGAGG CCCCA AACCA TGGGG ACCTC CATCC AACCA AACCA

K12b         AGGTG GACCA GGAGG CCCCA AACCA TGGGG ACCTC CATCC AACCA AACCA
K5a          AGGTG GACCA GGAGG CCCCA AACCA TGGGG ACCTC CATCC AACCA AACCA
K6a          AGGTG GACCA GGAGG CCCCA AACCA TGGGG ACCTC CATCC AACCA AACCA
K12a         AGGTG GACCA GGAGG CCCCA AACCA TGGGG ACCTC CATCC AACCA AACCA


pCog    601  CATCC ACGAC AACTG AGGCC TCAAC AAGCA CTTCC ACCAC GACAG CTTCG
                                    ------------------
K12b         CATCC ACGAC AACTG AGGCC TCAAC AAGCA CTTCC ACCAC GACAG CTTCG
K5a          CATCC ACGAC AACTG AGGCC TCAAC AAGCA CTTC
K6a          CATCC ACGAC AACTG AGGCC TCAAC AAGCA CTTCC ACCAC GACAG CTTCG
K12a         CATCC ACGAC AACTG AGGCC TCAAC AAGCA CTTCC ACCAC GACAG CTTCG
                                    ------------------

pCog    651  TCCAC TACAG TGTCC TCCAC TACAG AGTCC TCTAC GGAAT CCTCG ACGGA

K12b         TCCAC TACAG TGTCC TCCAC TACAG AGTCC TCTAC GGAAT CCTCG ACGGA
K6a          TCCAC TACAG TGTCC TCCAC TACAG AGTCC TCTAC GGAAT CCTCG ACGGA
K12a         TCCAC TACAG TGTCC TCCAC TACAG AGTCC TCTAC GGAAT CCTCG ACGGA


pCog    701  ATCAT CCACA GCATC CTCCA CAGAA TAAAT CGCTT GACAT GTTCC CCTTC

K12b         ATCAT CCACA GCATC CTCCA CAGAA TAAAT CGCTT GACAT GTTCC CCTTC
K6a          ATCAT CCACA GCATC CTCCA CAGAA TAAAT CGCTT GACAT GTTCC CCTTC
K12a         ATCAT CCACA GCATC CTCCA CAGAA TAAAT CGCTT GACAT GTTCC CCTTC


pCog    751  GAGTT TTGCC CATTC GGCTA TTCCG GGTTC ATAAA TAATT GTATA TATCA

K12b         GAGTT TTGCC CATTC GGCTA TTCCG GGTTC ATAAA TAATT GTATA TATCA
K6a          GAGTT TTGCC CATTC GGCTA TTCCG GGTTC ATAAA TAATT GTATA TATCA
K12a         GAGTT TTGCC CATTC GGCTA TTCCG GGTTC ATAAA TAATT GTATA TATCA


pCog    801  AAAAG CGAAT CTGTG CGTAATAAAT TTTTT TTTTT AGCTC AATCG TGATT

K12b         AAAAG CGAAT CTGTG CGTAA TAAAT TTTTT TTTTT AGCTC AAAAA AAAAA
K6a          AAAAG CGAAT CTGTG CGTAA TAAAT TTTTT TTTTT AGCTC AAAAA AAAAA
K12a         AAAAG CGAAT CTGTG CGTAA TAAAT TTTTT TTTTT AGCTC


pCog    851  TTTTT ATCTA AGTTA ATCCT TGAAT GCCAA ATAGT TATCC AAAGG ATCTT
pCog    901  CATCC ATGTT TGGCT TATGA TCAAC TTGAT GTATC AAGCC AGAAG GTACA
pCog    951  TGCGA GTACC ACGCG AGCCA TTTAA TTACT TCGGC CAATA CTAAA ACTTT
pCog   1001  TCGTG AGTTA AGTAA TTTTA GAACA ATTAG TTAAA TGGCA TTATG CAATG
pCog   1051  CCAGT GCTAA ATTAT CCACT GGCGA TTAGG AAGTG CATTG AGAAT TTGAA
pCog   1101  GCAAA TCATT CCCCG GGCCA ATTGG AATCG AACTT ATTCG AGCCG GCCTC
pCog   1151  GTTTC TCGCT TGGGT GAAAG CCAAA TTGTT TACCA ATTAA GTGCA TTTGG
pCog   1201  TCAAT TTGCG ACAGA ACTGG CGCTA GTGGA ATCCA AGAAG AAGGT ACAGA
pCog   1251  AGCT
```

Figure 5.4 Alignment of cDNAs and genomic DNA sequences

The sequences of the genomic fragment of *pCog* are compared to the sequences of the corresponding cDNAs K12a, K12b, K5a and K6a. Nucleotide reference numbers are given to the left. The first nucleotide number at the beginning of each cDNA is given with respect to the numbers of the genomic sequence. The sequences which are under- and over-lined at position 231 to 336 indicate the probe used in figure 5.3, those which are under- and over-lined at position 615 to 634 (----)indicate the sequencing primer used. Transcription elements, such as the TATA box at position 228, the 'Start' signal at position 258 and the poly adenylation signal at position 818, are highlighted in bold letters.

eliminates the third , which is from the opposite strand (Fig. 5.2 and 5.5). The open reading frame termed ORF II (Fig. 5.2, Fig. 5.5) extends from nucleotide 275 to 724, with the first methionine at the position 291, which is the sixth amino acid residue. This could be the starting methionine (Fig. 5.5). Thirty three nucleotides upstream to this ATG codon is the sequence 'ATCAGTT', which is the consensus sequence deduced for the start of transcription in *Drosophila* (Hultmark *et al.*, 1986). The distance from the first T of the TATA box to position +1 (of the start consensus) has been shown to be between 27 and 33 nucleotides (Hultmark *et al.*, 1986), which agrees well with the twenty nine nucleotides present between the start consensus sequence and the 'TATA box' in the *pCog* sequence.

In comparison the first open reading frame from nucleotide 163 to 835 shows the first methionine 30 amino acids after the stop codon (Fig. 5.5). This methionine is only 15 nucleotides downstream of the TATA box and upstream of the proposed consensus start site. Thus ORF I could not use the proposed consensus start site at nucleotide 258 and there is no other suitable sequence in the region. Additionally the start codon is only 20 nucleotides downstream of the first nucleotide of the proposed TATA box which is closer than usually found (Hultmark *et al.*, 1986).

The most significant evidence in favour of ORF II compared to ORF I is the manifestation of the $(GGX)_n$-repeat at the amino acid level. Although variations to the $(GGX)_n$-repeat occur at the nucleotide level, they are mostly in the third position of the triplets in ORF II (Table 5.1) and therefore rarely give rise to different amino acids. However, with ORF I, the nucleotide variations are at

Figure 5.5 Translation of the genomic sequence into open reading frames

```
        10         20         30         40         50         60         70
AGCTTTGTAACCAACAATTTTGGAAATAAACCTAAACATGAATCAACCATTGTCTACAAACTCATATTCT
 S  F  V  T  N  N  F  G  N  K  P  K  H  E  S  T  I  V  Y  K  L  I  F
  A  L  *  P  T  I  L  E  I  N  L  N  M  N  Q  P  L  S  T  N  S  Y  S
   L  C  N  Q  Q  F  W  K  *  T  *  T  *  I  N  H  C  L  Q  T  H  I  L

        80         90        100        110        120        130        140
TTCCATCTGACAGAACCCCAGAGTGAAGGCCAATGTCAATTGCGAAGTGACTGACTTGGAAAGTGGAATT
F  P  S  D  R  T  P  E  *  R  P  M  S  I  A  K  *  L  T  W  K  V  E  F
 F  H  L  T  E  P  Q  S  E  G  Q  C  Q  L  R  S  D  *  L  G  K  W  N
  S  I  *  Q  N  P  R  V  K  A  N  V  N  C  E  V  T  D  L  E  S  G  I

       150        160        170        180        190        200        210
TACAAATCCACACTGACGTTAGTTTGCTCAACGCGCAAACGTCACCGATGGAACACATGTGCCACCGAGA
  T  N  P  H  *  R  *  F  A  Q  R  A  N  V  T  D  G  T  H  V  P  P  R
L  Q  I  H  T  D  V  S  L  L  N  A  Q  T  S  P  M  E  H  M  C  H  R  D
 Y  K  S  T  L  T  L  V  C  S  T  R  K  R  H  R  W  N  T  C  A  T  E

       220        230        240        250        260        270        280
CTTTTAGCAGACGGTGGGTATATATACCGATGCCAAGTAATGCCGAGATCAGTTCGAGTCGTTAACTATC
 L  L  A  D  G  G  Y  I  Y  R  C  Q  V  M  P  R  S  V  R  V  V  N  Y
  F  *  Q  T  V  G  I  Y  T  D  A  K  *  C  R  D  Q  F  E  S  L  T  I
T  F  S  R  R  W  V  Y  I  P  M  P  S  N  A  E  I  S  S  S  R  *  L  S

       290        300        310        320        330        340        350
AGATAACGAAATGCGTACCCTTATCCTTGTTACTCTTGTTGCCCTGGTCGCAGTGGCCTCCGCCCAAGGA
Q  I  T  K  C  V  P  L  S  L  L  L  L  L  P  W  S  Q  W  P  P  P  K  D
 R  *  R  N  A  Y  P  Y  P  C  Y  S  C  C  P  G  R  S  G  L  R  P  R
  D  N  E  M  R  T  L  I  L  V  T  L  V  A  L  V  A  V  A  S  A  Q  G

       360        370        380        390        400        410        420
CCTGGTCCTTGGGGTCCCGGTGGACCTGGTGGACCTGGTGGACCAGGTCGTGGTCGCGGAGGACCAGGAC
 L  V  L  G  V  P  V  D  L  V  D  L  V  D  Q  V  V  V  A  E  D  Q  D
T  W  S  L  G  S  R  W  T  W  W  T  W  W  T  R  S  W  S  R  R  T  R  T
 P  G  P  W  G  P  G  G  P  G  G  P  G  G  P  G  R  G  R  G  G  P  G

       430        440        450        460        470        480        490
GTGGACCTGGCGGACCAGGTGGCCCAGGAGGTCGCGGACCTGGAGGACCTGGAGGACCAGGAGGCCCAGG
 V  D  L  A  D  Q  V  A  Q  E  V  A  D  L  E  D  L  E  D  Q  E  A  Q
  W  T  W  R  T  R  W  P  R  R  S  R  T  W  R  T  W  R  T  R  R  P  R
R  G  P  G  G  P  G  G  P  G  G  R  G  P  G  G  P  G  G  P  G  G  P  G

       500        510        520        530        540        550        560
AGGCCCAGGAGGACCAGGAGGACCAGGAGGCCCAGGAGGACCAGGATGCCCAGGAGGACCAGGTGGACCA
E  A  Q  E  D  Q  E  D  Q  E  A  Q  E  D  Q  D  A  Q  E  D  Q  V  D  Q
 R  P  R  R  T  R  R  T  R  R  P  R  R  T  R  M  P  R  R  T  R  W  T
  G  P  G  G  P  G  G  P  G  G  P  G  G  P  G  C  P  G  G  P  G  G  P

       570        580        590        600        610        620        630
GGAGGCCCCAAACCATGGGGACCTCCATCCAACCAAACCACATCCACGACAACTGAGGCCTCAACAAGCA
  E  A  P  N  H  G  D  L  H  P  T  K  P  H  P  R  Q  L  R  P  Q  Q  A
R  R  P  Q  T  M  G  T  S  I  Q  P  N  H  I  H  D  N  *  G  L  N  K  H
 G  G  P  K  P  W  G  P  P  S  N  Q  T  T  S  T  T  T  E  A  S  T  S

       640        650        660        670        680        690        700
CTTCCACCACGACAGCTTCGTCCACTACAGTGTCCTCCACTACAGAGTCCTCTACGGAATCCTCGACGGA
L  P  P  R  Q  L  R  P  L  Q  C  P  P  L  Q  S  P  L  R  N  P  R  R
  F  H  H  D  S  F  V  H  Y  S  V  L  H  Y  R  V  L  Y  G  I  L  D  G
T  S  T  T  T  A  S  S  T  T  V  S  S  T  T  E  S  S  T  E  S  S  T  E
```

Figure 5.5 continued

```
       710        720        730        740        750        760        770
ATCATCCACAGCATCCTCCACAGAATAAATCGCTTGACATGTTCCCCTTCGAGTTTTGCCCATTCGGCTA
N H P Q H P P Q N K S L D M F P F E F C P F G Y
 I I H S I L H R I N R L T C S P S S F A H S A
  S S T A S S T E * I A * H V P L R V L P I R L

       780        790        800        810        820        830        840
TTCCGGGTTCATAAATAATTGTATATATCAAAAAGCGAATCTGTGCGTAATAAATTTTTTTTTTTAGCTC
  S G F I N N C I Y Q K A N L C V I N F F F * L
I P G S * I I V Y I K K R I C A * * I F F F S S
 F R V H K * L Y I S K S E S V R N K F F F L A

       850        860        870        880        890        900        910
AATCGTGATTTTTTTATCTAAGTTAATCCTTGAATGCCAAATAGTTATCCAAAGGATCTTCATCCATGTT
 N R D F F I * V N P * M P N S Y P K D L H P C
  I V I F L S K L I L E C Q I V I Q R I F I H V
Q S * F F Y L S * S L N A K * L S K G S S S M F

       920        930        940        950        960        970        980
TGGCTTATGATCAACTTGATGTATCAAGCCAGAAGGTACATGCGAGTACCACGCGAGCCATTTAATTACT
L A Y D Q L D V S S Q K V H A S T T R A I * L L
 W L M I N L M Y Q A R R Y M R V P R E P F N Y
  G L * S T * C I K P E G T C E Y H A S H L I T

       990       1000       1010       1020       1030       1040       1050
TCGGCCAATACTAAAACTTTTCGTGAGTTAAGTAATTTTAGAACAATTAGTTAAATGGCATTATGCAATG
  R P I L K L F V S * V I L E Q L V K W H Y A M
F G Q Y * N F S * V K * F * N N * L N G I M Q C
 S A N T K T F R E L S N F R T I S * M A L C N

      1060       1070       1080       1090       1100       1110       1120
CCAGTGCTAAATTATCCACTGGCGATTAGGAAGTGCATTGAGAATTTGAAGCAAATCATTCCCCGGGCCA
 P V L N Y P L A I R K C I E N L K Q I I P R A
  Q C * I I H W R L G S A L R I * S K S F P G P
A S A K L S T G D * E V H * E F E A N H S P G Q

      1130       1140       1150       1160       1170       1180       1190
ATTGGAATCGAACTTATTCGAGCCGGCCTCGTTTCTCGCTTGGGTGAAAGCCAAATTGTTTACCAATTAA
N W N R T Y S S R P R F S L G * K P N C L P I K
 I G I E L I R A G L V S R L G E S Q I V Y Q L
  L E S N L F E P A S F L A W V K A K L F T N *

      1200       1210       1220       1230       1240       1250
GTGCATTTGGTCAATTTGCGACAGAACTGGCGCTAGTGGAATCCAAGAAGAAGGTACAGAAGCT
  C I W S I C D R T G A S G I Q E E G T E A
S A F G Q F A T E L A L V E S K K K V Q K L
 V H L V N L R Q N W R * W N P R R R Y R S F
```

Figure 5.5 illustrates the translation of the genomic sequence corresponding to the genomic fragment pCog into the predicted amino acid sequence. The two open reading frames termed ORF II (from nucleotide 275 to 724) and ORF I (from nucleotide 163 to 835) are underlined starting at ATG.

Note:
Analysis of preferred codon usage in *Drosophila* was carried out and did not discriminate between the three open reading frames.

the first nucleotide of each triplet and almost always result in a different amino acid. Thus if the (GGX) repeat is part of the coding region, it can be concluded that the second open reading frame is more likely to present the coding frame of the gene.

The sequence of the cDNAs corresponds to this interpretation. cDNA K5a begins at the potential start site. None of the cDNAs has sequences upstream of this start site.

5.3.3 The predicted protein corresponding to the "collagen-like" gene

The nucleotide sequence of the "collagen-like" gene was translated into the predicted amino acid sequence and compared to sequences in the following DATAbases, the NBRF protein bank (release 29.0), the Swiss protein bank (release 14.0) and the Genpept library (release 63.0) to search for similar sequences using the FASTA algorithm (Pearson & Lipman, 1988). The searches were done using the Wisconsin Package (version 7.0, April 1991; copyright 1991 by John Devereux; Devereux *et al.*, 1984). The following terminology is used: $(GGX)_n$ repeat refers to the nucleotide repeat illustrated in Table 5.1. Because the translation of the triplets (GGX) with X being any nucleotide codes for glycine, and (CCX) codes for proline, the translation of the $(GGX)_n$ repeat results in a glycine-glycine-proline repeat hereafter abreviated as gly gly pro.

Although the evidence points towards ORF II as the coding frame, ORF I, ORF II and also ORF III were used for translation and the resulting amino acid sequences were used as query sequences for database searches.

The predicted protein translated from the first open reading frame showed similarities of approximately 20% (in 100 amino acids) to various proteins (Table 5.3), e.g. the androgen receptor in

Table 5.3

similar gene	% ID	aa	N	references
circumsporozoite protein precursor, Plasmodium b.	37.0	140	230	Lal *et al.*, 1987
Collagen alpha 1(III) chain, bovine	38.5	117	207	Fietzek *et al.*, 1979
probable nuclear antigen, Epstein Barr virus	33.1	133	201	Baer *et al.*, 1984
collagen 2, Caenorhabditis elegans	42.3	104	198	Kramer *et al.*, 1982
collagen 1, Caenorhabditis elegans	40.6	106	197	Kramer *et al.*, 1982
procollagen alpha 1(III) chain precursor, human	35.7	140	192	Seyer and Kang, 1977
collagen alpha 1(I) chain, chicken	33.3	123	183	Highberger *et al.*, 1982
procollagen alpha 1(I) chain precursor, chicken	33.3	123	183	Highberger *et al.*, 1982
collagen alpha 1(I) chain, mouse	35.0	117	182	French *et al.*, 1985
collagen alpha 1(I) chain, bovine	33.3	120	182	Rautenberg *et al.*, 1972

Table 5.3 List of nucleotide sequences with the 10 highest scores retrieved from the NBRF databank. The similarities of the predicted *pCog* protein using ORF II to the best scoring proteins (column 1) are listed in percent identity (% ID in column 2) within an amino acid stretch (aa in column 3) and ranked by the optimal score (N) in column 4.

humans (21.8%) (Chang *et al.*, 1988; Luhbahn *et al.*, 1988), the ubiquinol cytochrome reductase-c protein in yeast (19.5%) (Van Loon *et al.*, 1984), and the AAC3 protein in *Dictyostelium* (19.8%)

(Shaw *et al.*, 1989). In each case the similarity found was due to a run of either glutamine (Q) or aspartic acid (D), which is obviously not of significance (data not shown).

A similar result is obtained when the amino acid sequence used as the query sequence is translated from ORF II. Initially the similarity appeared to be promising, for example 40% in an 104 amino acid overlap to a collagen of *Caenorhabditis elegans* (Fig. 5.7; Table 5.3). However, closer examination shows that the similarities are most likely not significant.

One of the special features of a collagen molecule is the amino acid sequence "gly-X-Y", where X and Y represent any two non-equivalent amino acids (Ramachandran, 1967; Fessler and Fessler, 1978). In 60% of the cases X represents a proline. The translation of the $(GGX)_n$ repeat (Table 5.1) can be seen as "gly, pro, gly" starting with the second triplet in table 5.1, which results in a relatively high degree of similarity between the predicted *pCog* protein and the collagens (Table 5.3). Nine out of the ten sequences most similar to the "collagen-like" protein are collagens (Table 5.3). The similarities range from 33 to 40%. However, these similarities are most likely not significant for the following reasons.

Firstly, the homology is only present within the $(GGX)_n$ repeat. No similarities are found outside this repeat (Fig. 5.7). Secondly, although the glycine and proline content is very high in

Figure 5.7 Alignment of the pCog amino acid sequence predicted using ORF II to similar peptides retrieved from various databases

```
protein 1: collagen 2, Caenorhabditis elegans

SCORES 198, 42.3% identity in 104 aa overlap

                              10        20        30        40
Pcog.G                LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPGRGRGGPG--
                                        |:: |||| || ||:||:| : ::|:||
col 2  CQPCPGGPPGPAGPAGPPGPPGPDGNPGSPAGPSGPGPAGPPGPAGPAGNDGAPGAPGGP
        150       160       170       180       190       200

         50        60        70        80        90       100
Pcog.G --RGPGGPGGPGGRGPGGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPKPWGPPSN
         :|::::||||::||:|| ||:||:| :|: |:|||| :|  | ||::|:| : | |::
col 2  GEPGASEQGGPGEPGPAGPPGPAGPAGNDGAPGTGGPGPAGPKGPPGAAGAPGADGNPGG
        210       220       230       240       250       260

        110       120       130       140       150
Pcog.G QTTSTTTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       ::|:::::::::::
col 2  PGTAGKPGGEGEKGICPKYCAIDGGVFFEDGTRRR
        270       280       290       300


protein 2: sporozoite surface antigen, Plasmodium berghei yoelli

SCORES 230, 37.1% identity in 140 aa overlap

                                  10        20        30        40
Pcog.G                      LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPGRG
                                        | : |:::::  ::|| :| |||:| |||
prot 2 KKDDLPKEEKKDDPPKDPKKDDPPKEAQNKLNQPVVADENVDQGPGAPQGPGAPQGPGAP
        100       110       120       130       140       150

              50        60        70        80        90       100
Pcog.G RG-GPGRGPGGPGGPGG-RGPGGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGP-KP
       :| |: :|||:| |||: :|||:| |||:| |||:| |||:| |||:| |||  | |||:| ||  |
prot 2 QGPGAPQGPGAPQGPGAPQGPGAPQGPGAPQGPGAPQGPGAPQGPGAPQGPGAP
        160       170       180       190       200       210

             110       120       130       140       150
Pcog.G WGPPSNQTTSTTTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       || : |:::::  |::  ::  ::  ::  :: ::: ::: :::::::::
prot 2 QGPGAPQGPGAPQEPPQQPPQQPPQQPPQQPPQQPPQQPPQQPRPQPDGNNNNNNNNGNN
        220       230       240       250       260       270

prot 2 NEDSYVPSAEQILEFVKQISSQLTEEWSQCSVTCGSGVRVRKRKNVNKQPENLTLEDIDT
        280       290       300       310       320       330


protein 3: collagen alpha I(III) chain, bovine

SCORES 207, 38.5% identity in 117 aa overlap

                     10        20        30        40        50
Pcog.G         LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPGRGRGGPGRGPGGP
                                        :||  | ||| |::||: |:| :|: || :
col    ERGEQGPPGPAGFPGAPGQNGEPGAKGERGAPGEKGEGGPPGAAGPAGGSGPAGP-PGPQ
            650       660       670       680       690       700

          60        70        80        90       100       110
Pcog.G GGPGGRG-PGGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPKPWGPPSNQTTSTTT
       |  |:|| |||||::| |||:| || ||: |:|| ||::|:||   | |||:::::::::
col    GVKGERGSPGGPGAAGFPGGRGPPGPPGSNGNPGPPGSSGAPGKDGPPGPPGSNGAPGSP
             710       720       730       740       750       760

          120       130       140       150
Pcog.G EASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       : |:::::::::::  ::::::::::::
col    GISGPKGDSGPPGERGAPGPQGPPGAPGPLGIAGLTGARGLAGPPGMPGARGSPGPQGIK
             770       780       790       800       810       820
```

```
protein 4: probable nuclear antigen, Epstein-Barr virus,

SCORES 201, 33.1% identity in 133 aa overlap

                          10        20        30        40
Pcog.G              LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPGRGRG-GP
                                    :::::::|:|  |:|||:||:||:| | | |:
nuc     QKRPSCIGCKGTHGGTGAGAGAGGAGAGGAGAGGGAGAGG-GAGGAGGAGGAGAGGGAGA
          80        90        100       110       120       130

         50        60        70        80        90        100
Pcog.G GRGPGGPGGPG-GRGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPKPWGPPSN
       | |:|||:||:| | |:|: ||:|||:|: ||:|||:|||:|: | :|: ||:|||: : | :::
nuc     GGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGGAGGAGAGGGAGAGGGAGGAGAGGGAGG
          140       150       160       170       180       190

         110       120       130       140       150
Pcog.G QTTSTTTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       ::::::::::::::::::|:::::::::::::::::::::::|:::
nuc     AGGAGAGGGAGAGGAGGAGGAGAGGAGAGGGAGGAGGAGAGGAGAGGAGAGGAGAGG
          200       210       220       230       240       250

nuc     AGAGGAGGAGAGGAGGAGAGGGAGGAGAGGGAGGAGAGGAGGAGAGGAGGAGAGGAGGAG
          260       270       280       290       300       310


protein 5: collagen 1, Caenorhabditis elegans

SCORES 197, 40.6% identity in 106 aa overlap

                        10        20        30        40        50
Pcog.G            LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPGRGRGGPGRG
                                    :::::||: || || ||:| : |  ||: :
col 1   CKPCPQGPPGPPGPPGAPGDPGEAGTPGRPGTDAAPGSPGPRGPPGPAGEA-GAPGPAGE
              150       160       170       180       190       200

               60        70        80        90        100       110
Pcog.G PGGPGGPGGRGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPKPWGPPSNQTTS
       ||:|: ::  :||:||:||::| || || ||:||: | || ||  |:| | |||:::: |
col 1   PGTPAISEPLTPGAPGEPGDSGPPGPPGPPGAPGNDGPPGPPGPKGAPGPDGPPGADGQS
              210       220       230       240       250       260

              120       130       140       150
Pcog.G TTTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       ::::::::::|:::::
col 1   GPPGPPGPAGTPGEKGICPKYCALDGGVFFEDGTRR
              270       280       290


protein 6,: procollagen alpha I(III) chain precursor, Human

SCORES 192, 35.7% identity in 140 aa overlap

                            10        20        30          40
Pcog.G                LSDNEMRTLILVTLVALVAVASAQG-PGPWGPGGPGG-PG--GPG
                                    :| ::|:::| || :||:||:| ||  ||:
procol GERGEAGIPGVPGAKGEDGKDGSPGEPGANGLPGAAGERGAPGFRGPAGPNGIPGEKGPA
        280       290       300       310       320       330

               50        60        70        80        90
Pcog.G RGRGGPGR-GPGGPGGPGGRG--PGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGP
       :||:||: || |::|  ||:  ||||| :| ||:|||||: | || ||::|::| || |
procol GERGAPGPAGPRGAAGEPGRDGVPGGPGMRGMPGSPGGPGSDGKPGPPGSQGESGRPGPP
        340       350       360       370       380       390

      100       110       120       130       140       150
Pcog.G KPWGPPSNQTTSTTTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
       | ||:::::: : :::::::::::: :::  :::: :: ::::::::::::
procol GPSGPRGQPGVMGFPGPKGNDGAPGKNGERGGPGGPGPQGPPGKNGETGPQGPPGPTGPG
        400       410       420       430       440       450

procol GDKGDTGPPGPQGLQGLPGTGGPPGENGKPGEPGPKGDAGAPGAPGGKGNAGAPGERGPP
        460       470       480       490       500       510


protein 7: collagen alpha I(I) chain, chicken

SCORES  183, 33.3% identity in 123 aa overlap
```

```
                  10        20        30        40        50
Pcog.G       LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGP--GRGRGGPGRGPG
                                      :||: |: |::|::||  :|| :||:  ||
col 1  PSGASGERGPPGPMGPPGLAGPPGEAGREGAPGAEGAPGRDGAAGPKGDRGETGPAGPPG
          830       840       850       860       870       880

           60        70        80        90        100       110
Pcog.G GPGGPGGRGPGGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPK-PWGPPSNQTTST
       :||:||::|| ||:| :|  | :||:||:|| ||: : ||:|| ||:   | :::|::::
col 1  APGAPGAPGPVGPAGKNGDRGETGPAGPAGPPGPAGARGPAGPQGPRGDKGETGEQGDRG
          890       900       910       920       930       940

           120       130       140       150
Pcog.G TTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
        :: :: ::  :::::: ::::::::::|: ::
col 1  MKGHRGFSGLQGPPGPPGAPGEQGPSGASGPAGPRGPPGSAGAAGKDGLNGLPGPIGPPG
          950       960       970       980       990      1000

col 1  PRGRTGEVGPVGPPGPPGPPGPPGPPSGGFDLDGGRYY
         1010      1020      1030      1040


protein 8: collagen alpha I(I) chain chicken

SCORES 183, 33.3% identity in 123 aa overlap

                  10        20        30        40        50
Pcog.G       LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGP--GRGRGGPGRGPG
                                      :||: |: |::|::||  :|| :||:  ||
col 1  PSGASGERGPPGPMGPPGLAGPPGEAGREGAPGAEGAPGRDGAAGPKGDRGETGPAGPPG
          830       840       850       860       870       880

           60        70        80        90        100       110
Pcog.G GPGGPGGRGPGGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPK-PWGPPSNQTTST
       :||:||::|| ||:| :|  | :||:||:|| ||: : ||:|| ||:   | :::|::::
col 1  APGAPGAPGPVGPAGKNGDRGETGPAGPAGPPGPAGARGPAGPQGPRGDKGETGEQGDRG
          890       900       910       920       930       940

           120       130       140       150
Pcog.G TTEASTSTSTTTASSTTVSSTTESSTESSTESSTASSTE
        :: :: ::  :::::: ::::::::::|: ::
col 1  MKGHRGFSGLQGPPGPPGAPGEQGPSGASGPAGPRGPPGSAGAAGKDGLNGLPGPIGPPG
          950       960       970       980       990      1000

col 1  PRGRTGEVGPVGPPGPPGPPGPPGPPSGGFDLDGGRYY
         1010      1020      1030      1040


protein 9: collagen alpha I(I) chain, mouse

SCORES 182, 35.0% identity in 117 aa overlap

                10        20        30        40        50
Pcog.G  LSDNEMRTLILVTLVALVAVASAQGPGPWGPGGPGGPGGPG-RGRGGPGRGPGGPGGPG
                                      :||  |:||: | || :||:  ||:||:||
col 1  ERGPPGPMGPPGLAGPPGESGREGSPGAEGSPGRDGAPGAKGDRGETGPAGPPGAPGAPG
          470       480       490       500       510       520

       60        70        80        90        100       110
Pcog.G GRGPGGPGGPGGPGGPGGPGGPGGPGCPGGPGGPGGPK-PWGPPSNQTTSTTTEAST
       ::|| ||:| :|  | :||:||:|| ||: : ||:|| ||:   | :::|:::::::: ::
col 1  APGPVGPAGKNGDRGETGPAGPAGPIGPAGARGPAGPQGPRGDKGETGEQGDRGIKGHRG
          530       540       550       560       570       580

       120       130       140       150
Pcog.G STSTTTASSTTVSSTTESSTESSTESSTASSTE
        ::  :::::: |::::::::::|: ::
col 1  FSGLQGPPGSPGSPGEQGPSGASGPAG
          590       600       610
```

Figure 5.7 Amino acid alignment of the "collagen-like" protein using ORF II, with some representative proteins, retrieved from the SWISS-PROT, NBRF/prot and GenPept libraries (Table 5.3). Identical amino acids are marked using bars, conservative exchanged amino acids are marked with double dots.

collagens, they rarely contain more than one glycine per amino acid triplet, an attribute which, in contrast, is the rule in the translated $(GGX)_n$ repeat of the "collagen-like" protein. In fact the highest score was obtained with ORF III (data not shown) because the translation of the $(GGX)_n$ repeat on the opposite strand results in the amino acid repeat "glycine, proline, proline", which indeed resembles the "gly-X-Y" of collagens.

A third reason why the similarity of the "collagen-like" protein to collagens can only be regarded as superficial lies within the sequence of the $(GGX)_n$ repeat. Unlike the helical domain of collagens, the amino acid translation of the $(GGX)_n$ repeat fails to sustain the rhythm of gly-X-Y (Fig. 5.8). Three changes would be necessary to sustain the repeat of gly-pro-gly in the amino acid translation of $(GGX)_n$ - the deletion of a residue at position 37 and one at position 44, and the addition of a residue at position 57 (Fig. 5.8). The regular occurence of a glycine residue at every third position within the helical domain is the main feature which has been used to define collagens (Adams 1978). Therefore the conclusion can be drawn that the "collagen-like" protein is not a collagen.

A closer examination of the first eighteen amino acids in ORF II revealed a noteworthy feature of the "collagen-like" gene as these amino acids formed the characteristics of a signal peptide (Fig. 5.8). One of the main characteristics of a signal peptide, as defined by Perlman and Halvorson (1983), is that the region of the first twenty amino acids has an overall hydrophobic composition with an uninterrupted hydrophobic "core" of 16 amino acids. In the predicted pCog protein using the translation from ORF II and

Figure 5.8 The predicted pCog protein

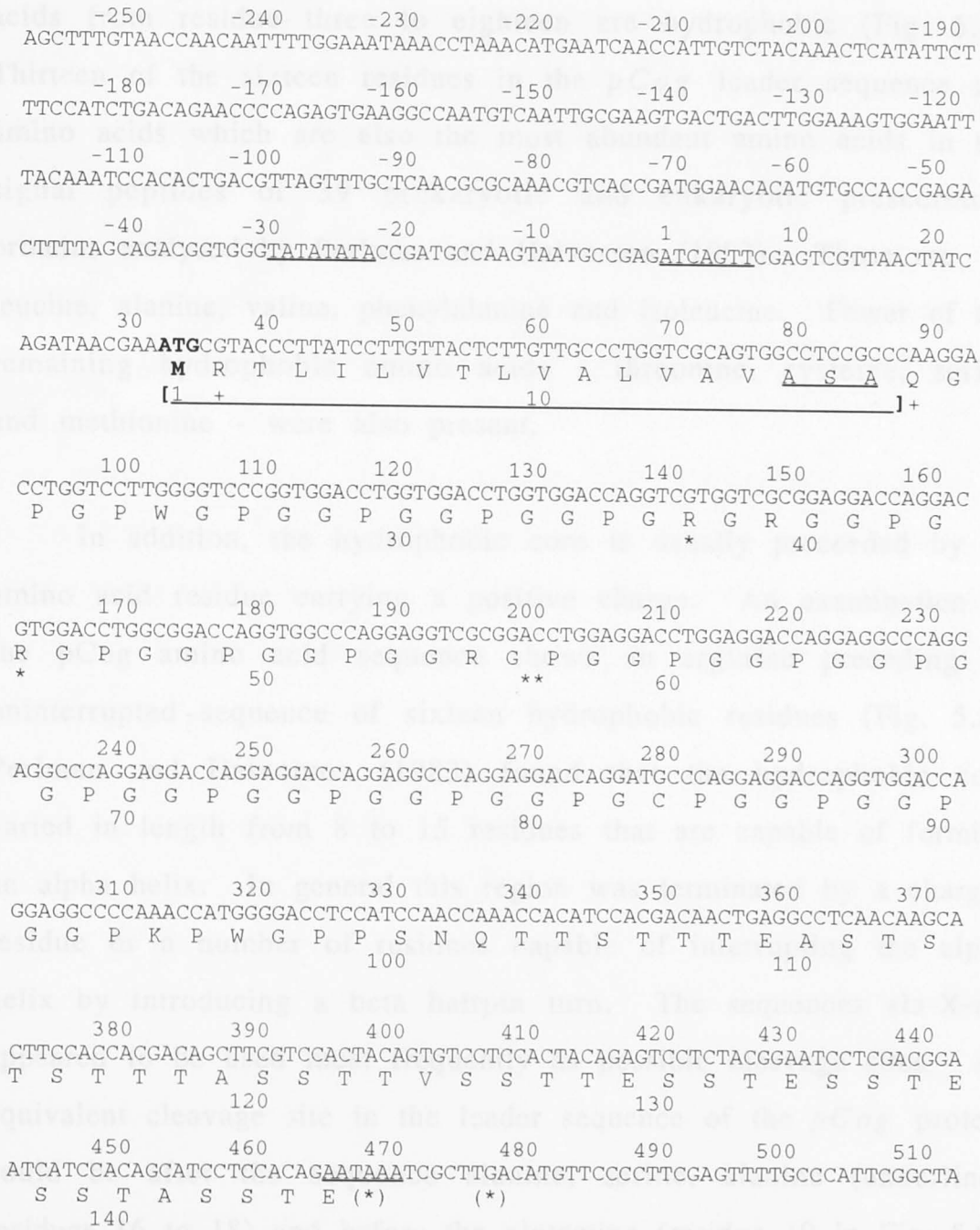


Figure 5.8 illustrates the predicted amino acid sequence corresponding to the genomic fragment pCog using ORF II. The amino acid sequence is in one letter code. The signal peptide from amino acid residue 1 to 18 is indicated in brackets, [__]. The possible cleavage sequence ala- X -ala (residue 16 to 18) is underlined, the positively charged residue arginine (R) and the polar residue glutamine (Q) are marked with a plus +. Stars at amino acid residues 37, 44, and 56/57 indicate variations to the repeat gly-X-Y.
The TATA box at position -29 and the polyadenylation signals are underlined. The proposed start signal is set as position 1. The start codon ATG and the first methionine are in bold letters.
The stop codons at position 470 and 479 are indicated (*).

counting from methionine as amino acid number one, 16 amino acids from residue three to eighteen are hydrophobic (Fig. 5.8). Thirteen of the sixteen residues in the *pCog* leader sequence are amino acids which are also the most abundant amino acids in the signal peptides of 39 prokaryotic and eukaryotic presecretory proteins analysed by Perlman and Halvorson (1983). These are leucine, alanine, valine, phenylalanine and isoleucine. Fewer of the remaining hydrophobic amino acids - threonine, cysteine, serine and methionine - were also present.

In addition, the hydrophobic core is usually preceeded by an amino acid residue carrying a positive charge. An examination of the pCog amino acid sequence shows an arginine preceding an uninterrupted sequence of sixteen hydrophobic residues (Fig. 5.8). Perlman and Halvorson (1983) found that the hydrophobic core varied in length from 8 to 15 residues that are capable of forming an alpha helix. In general this region was terminated by a charged residue or a number of residues capable of interrupting the alpha helix by introducing a beta hairpin turn. The sequences ala-X-ala appeared to be used most frequently as possible cleavage sites. An equivalent cleavage site in the leader sequence of the *pCog* protein could be after the sequence alanine, serine, alanine (underlined residues 16 to 18) and before the glutamine (residue 19 in Fig. 5.8). Although glutamine (Q) is not a charged residue, it is polar and as such not part of the leader sequence.

Since searches with the entire amino acid sequence failed to reveal any significant similarities, the predicted amino acid sequence using ORF II was divided into three sub-sequences to search individually for sequence similarities in the same data

bases. Sub-sequence one extends from the first amino acid to the 24th residue i.e. includes the leader signal peptide; sub-sequence two is the amino acid translation of the $(GGX)_n$ repeat, and sub-sequence three is the remaining sequence from residue 94 to residue 146, which includes the second nucleotide repeat (Table 5.2). However, apart from the leader peptide sequence, no significant homology was found when any of the sub-sequences was used for the searches.

It is worth noting that none of the searches carried out including a search using the $(GGX)_n$ repeat only, revealed any similarities to other genes with known helical structures as for example the acetylcholine receptor gene.

5.4 Discussion

The sequences of the cDNAs and genomic DNA corresponding to the position 0 to 1.3 on the molecular map (chapter 3, Fig. 3.1) were obtained in an attempt to elucidate the function of the "collagen-like" gene.

Three approaches were taken to define the coding region of the "collagen-like" gene; (i) comparisons of genomic and cDNA sequences; (ii) establishing long open reading frames (ORFs) and (iii) searches for translation initiation signals.

One of the difficulties in predicting the correct coding frame was that ORF I (Fig. 5.2) is slightly longer than ORF II. However, there are several lines of evidence which strongly suggest that ORF II is the frame used for translation. The most 5' extending cDNA was cDNA 5a which started at the start consensus site AT CAG TT (position 258, Fig. 5.4).

Another argument in favour of ORF II is that, using this coding frame, the first eighteen amino acids show the properties of a signal leader sequence, required by proteins which have to pass from the endoplasmatic reticulum to the extracellular matrix. It would be surprising if this feature was merely due to chance rather than to a specific function.

A third argument in favour of ORF II is that variations from the sequence motif $(GGX)_n$ occur in the third position of the coding

triplets. If the protein has a function, then there will be selective pressure to preserve the sequence of amino acids. Assuming that the present nucleotide sequence of the $(GGX)_n$ repeat is derived originally from the pure sequence form (GGX) (GGX) (CCX), one can argue that variations to this sequence motif $(GGX)_n$ are more likely to occur at the third nucleotide of the coding triplets. Using ORF I the same variations appear at the first position of the coding triplets, leading to the substitution of different amino acids.

Analysis of the predicted amino acid sequence of the "collagen like" gene.

The translation of the *pCog* sequence using ORF II reveals a fairly unusual protein in that more than half of the amino acid sequence is repetitive. The total length of the predicted pCog protein is 145 amino acids, starting with the methionine at nucleotide position 291 (Fig. 5.8). The first repeat extends from the 25th to the 93rd amino acid residue (Table 5.1, Fig. 5.8); the second repeat from residue 113 to residue 145 (Table 5.2). This means that out of 145 amino acids a total of 100 residues, or 68%, are part of a repetitive sequence.

A comparison of the sequence data to the sequences of known genes suggests that the "collagen-like" gene codes for a novel protein. Although there are some similarities to collagens (Table 5.3; Fig. 5.7), these are most likely not significant because they are based solely on the similarity of the sequence repeat termed $(GGX)_n$ (Table 5.1) to the triple helical domain common to collagens. The significance of the similarity to collagens is further lessened by the fact that changes within the amino acid sequence of the "collagen-like" protein occur to interrupt the rhythm of gly-X-Y. The regular appearance of the amino acid glycine in every third

position (gly-X-Y repeat) is the central characteristic of collagens since it is the basis for the triple helical structure which forms the rod-like molecule. The length of the gly-X-Y repeat also shows a difference to collagen molecules which are in general approximately 1000 amino acids long; of these up to 90% can be triple helical (Bentz *et al.*, 1983; Stacey *et al.*, 1988)

The function of the "collagen-like" gene remains speculative. It can be assumed from the Northern analysis (chapter 3, Fig. 3.1) that the gene is mainly transcribed in adult tissue and that it might have an extracellular function because of the leader peptide sequence. To gain further insight into the role of this gene one could ask where and when in the organism it is expressed. The initial information I obtained from the isolation of RNA from different developmental stages could be extended by labelling cells by *in situ* hybridization. Thus, using labelled probes to mark cells in either whole embryos or tissue sections, it would be possible to determine where the "collagen-like" gene is expressed. This approach has been successfully used to describe a number of genes in *Drosophila* which are important in development (Ingham, 1988; Pankratz and Jaeckle, 1990) and cell division (Edgar and O'Farrell, 1989; Glover, 1991).

Although it is clear from the sequence comparisons that the "collagen-like" protein does not belong to the collagens, it would be intriguing to analyse whether the high content of glycines and prolines of the "collagen-like" gene has a structural effect similar to that of the triple helix in collagens.

To further investigate the function of the "collagen-like" gene it is also necessary to test whether the "collagen-like" gene is

indeed part of the *LB20* locus. This question will be raised in the following chapter.

Chapter 6

Transformation of the constructs representing the genes *A112* and *LB20*

6.1 Introduction

Transposable elements are DNA segments which, as discrete units, are capable of changing their position within the genome of an organism (Engels 1989; Rio, 1991). Several classes of transposable elements have been identified in the genome of *Drosophila* (Spradling and Rubin, 1981). *Drosophila* germline gene transfer is based on the special behaviour of one family of transposable elements known as P-elements.

P-elements are able to transpose into the chromosomes of germline cells, if introduced prior to the time of pole cell formation (Rubin and Spradling 1982; Spradling and Rubin 1982). Three features of P-elements are of importance. Firstly, encoded in the four open reading frames of the 2.9kb P-element is the transposase - a site specific DNA-binding protein (Rio, 1991). Secondly, the sequence of the 2.9kb P-element showed that it contained a 31bp indirect repeat on either end (O'Hare and Rubin, 1983). These 31bp repeats are recognized by the transposase and define the borders of the sequence that will be integrated into the chromosome. The third important characteristic is the tissue-specificity. Although P-elements are transcribed in both germ line and somatic tissue, the transposition of P-elements occurs only in the germ line (Laski *et al.* 1986; Laski and Rubin 1989; Siebel and Rio, 1990). These attributes provide the basis for successful gene transfer into the *Drosophila* genome under experimental conditions.

Germ line transformation has been highly successful in the specific rescue of mutant alleles by wild-type DNA and is an important tool in linking biological function to specific molecular structures. Germ line transformation has been used to unequivocally determine which parts of a DNA sequence are required for gene function (Goldberg *et al.*, 1983; Spradling and Rubin, 1983; Henkemeyer *et al.*, 1987; Royden *et al.*, 1987).

In order to obtain stable integration, a transposition-defective vector is usually injected into embryos together with a helper P-element which is defective at one of the 31bp repeats but supplies the transposase (Spradling, 1986). The transposition-defective vector contains the intact repeats, the gene to be introduced and a marker gene; the helper P-element supplies the transposase function without being able to integrate itself.

Germ line transformation was clearly the method of choice to investigate which DNA sequences are actually necessary for the function provided by the complementation groups *A112* and *LB20*. If the correct DNA fragments could be successfully introduced into the germline of flies, it would be possible to cross the transgenic flies to flies bearing mutations at the *A112* or *LB20* loci. If the wild-type DNA fragment introduced contained all parts of the gene necessary for its function, then it would complement the function lost at the mutated locus. This chapter describes transformation experiments investigating the *A112* and *LB20* transcription units.

6.2 Materials and Methods

6.2.1 Cloning

The procedures used to clone the constructs have been described in chapters 2.2 and 4.2.

6.2.2 *Drosophila* stocks and media

All *Drosophila* strains used were obtained from Dr. G.L.G. Miklos (Research School of Biological Sciences, Australian National University, Canberra). Flies were reared at 20°C on standard maize meal, molasses, yeast medium (10g Bacto-agar, 15g sucrose, 40g yeast, 40g malt, 30ml Karo corn syrup and 10g soya flour were boiled in 1 l water for 20 minutes; the medium was cooled to 50°C before 4.5ml of propionic acid and 9ml of a 10% Nipagen solution in ethanol were added). Egg collections were made on apple juice plates (22.5g Bacto-agar was dissolved in 750ml water by boiling. 1.25g Nipagen and 22g sucrose were dissolved in 250ml of apple juice by boiling. The two solution were mixed and poured into 90mm petri dishes.) smeared with a live yeast suspension.

6.2.3 *Drosophila* transformation

Germ-line transformation of *Drosophila* was carried out essentially as described by Rubin and Spradling (1982) with the fly strain w^1 as host strain. The helper plasmid used to transform with the constructs was pUC hs $\pi\Delta$ 2-3 (Mullins *et al.*, 1989; a kind gift of Dr. Steven Delaney).

Eggs were collected for 1 hour, then transferred onto adhesive tape and manually dechorionated. Approximately 25 embryos at a time were transferred to a double-sided adhesive tape attached to a coverslip and incubated 1-5 minutes at room temperature (18°C) to desiccate before they were covered with oil. The embryos were injected with a mixture of recombinant and helper plasmids using needles pulled from microcapillaries containing an internal filament.

Injected embryos were kept under oil for 2 days, the surviving larvae were transferred to yeast vials and those which pupated and eclosed (G_0 adults) were mated to several flies of the w^1 host strain. After 2 days sterile flies were discarded and the fertile crosses were pooled, each pool containing 3 to 4 G_0 adults. Any transformed progeny (G_1 adults) were mated again to the host strain and their offspring were self-crossed in order to establish transformed "G_2 lines".

G_2 adults were screened in order to establish the segregation of the w^+ phenotype with respect to the sex chromosomes. In the case of autosomal segregation w^+ males were used for complementation tests.

6.3 Results

6.3.1 Transposon construction

Several P-element transformation vectors containing appropriate marker genes for identifying transformed flies have been constructed (Spradling 1986). The vector chosen was *pW8*, which contains the coding portion of the wild-type *white* gene fused to the heat shock protein 70 (hsp 70) gene promoter (Klemenz *et al.*, 1987). This vector was selected because it has several advantages. Firstly, white mutations are available which have a very strong phenotype, for example homozygous w^1 mutants have bleach white eyes. When successfully transformed with the *pW8* vector the wild-type w^+ white gene produces wild-type eye colour in the w^1 flies. The transformants can thus be easily recognized as red-eyed flies amongst white-eyed non-transformants. Secondly, the P-element vector *pW8* contains 12 unique restriction sites in the polylinker, enabling a wide variety of fragments with different ends to be introduced. Thirdly, the *pW8* vector is relatively small (8kb) which is important since the overall transposon size affects the transformation efficiency (Spradling, 1986).

A number of factors determine the choice of the insert. The DNA fragment to be inserted should contain only one functional transcription unit. Also, it is necessary to include all possible regulatory sequences since the introduced gene would be under the control of its own promotor. Further, the overall transposon size

should not be too large, since as mentioned above, the size of the transposon affects the transformation efficiency.

The *A112* construct:

In chapter 2 it was shown that at least part of the *A112* locus must reside in the overlap of *Df(1)JC77* and *Df(1)HM44*. On the physical map this overlap occupies the 1.2kb area of DNA from -11.2 to -10 (Fig. 2.13). Both deficiencies genetically disrupt the *A112* locus, however, it is quite likely that the *A112* transcription unit extends further in both directions.

One transcript of 2.2kb has been detected which corresponds to this DNA region and this transcript can be detected with probes originating from the -13 to -4.8 region (Fig. 2.14).

The sequence data (chapter 4) revealed that this transcription unit extends 550bp distal and at least 1700bp proximal to the *Eco* RI site at -10 and is therefore interrupted by both deficiencies. This transcript therefore appeared as a likely candidate for the *A112* gene. It was also deduced from the sequence data that the direction of transcription was orientated from proximal to distal.

In order to find convenient restriction sites for the construction of the transformation plasmid a detailed restriction map of the relevant DNA region was made (Fig. 6.1). A 12 kb *Bgl* II fragment from -14.6 to - 2.7 was chosen as a suitable fragment to be cloned into the *pW8* transformation vector (Fig.6.1, 6.2). This genomic fragment contains the 2.2kb transcription unit, sequences distal and proximal to it and, distally, part of an additional transcription unit which has been identified as the *sluggish* gene

(David Hayward, personal communication). Since this distal transcription unit is interrupted by the distal *Bgl* II site, it is unlikely to be functional and will not interfere in the subsequent experiments, yet it serves as a buffer zone between the transcript and sequences adjacent to the insertion.

Figure 6.1

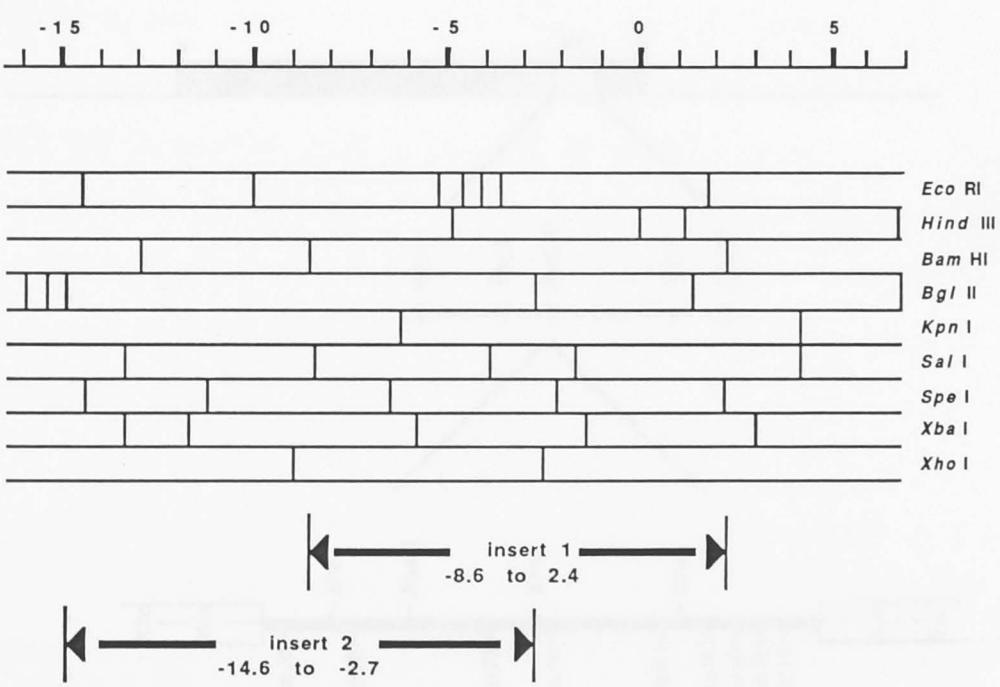


Fig.6.1 Restriction enzyme map of the *LB20/A112* area. The genomic DNA area containing the loci *LB20* and *A112* was mapped with several enzymes as indicated to the right (*Eco* RI and *Hind* III: Miklos pers. com.). All restriction sites are indicated as vertical bars. The co-ordinate system above corresponds to Fig. 2.1. Arrows indicate the limits of the genomic fragments used to construct the transposons. Insert 1 is contained within the transposon p[(w)LB20], insert 2 in transposon p[(w)A112].

The 12 kb *Bgl* II fragment (from position -14.6 to -2.7; insert 2 in Fig. 6.1) was excised from a recombinant λ bacteriophage and ligated into the *Bam* HI site of the polylinker of *pW8*. The resulting

construct (Fig. 6.2) is designated P[(w)A112], according to Spradling (1986), where 'P' indicates that the transposon is a defective P-element, w describes the marker gene and A112 is used as a synonym for the introduced DNA sequences. The overall size of P[(w)A112] is 20kb.

Figure 6.2 Design of the *A112* transformation construct

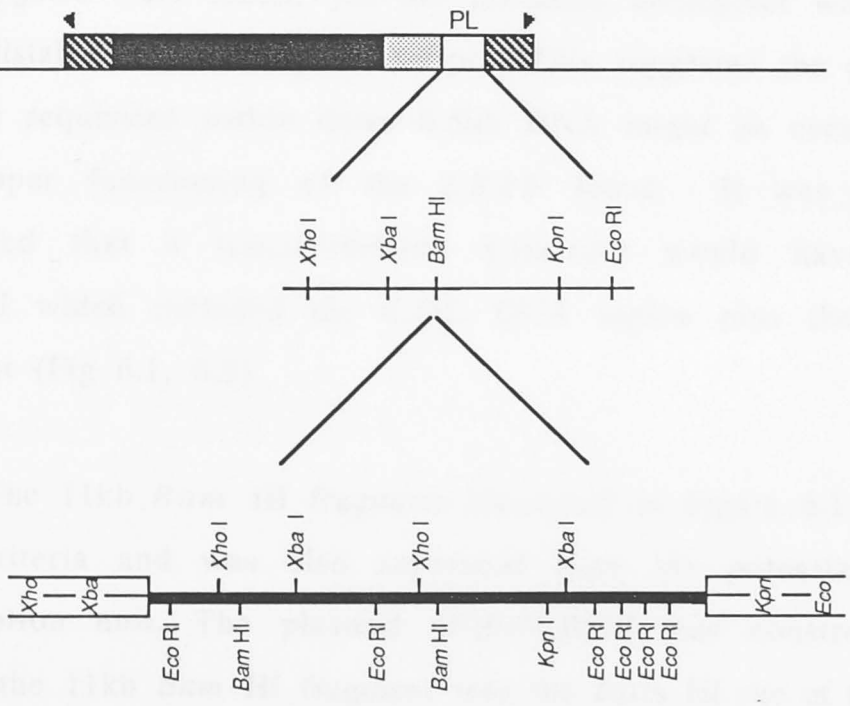


Fig. 6.2 Design of the transposon P[(w)A112]. The *pW8* vector is illustrated at the top. The *white* gene is indicated as filled box and the *hsp 70* gene as dotted box. The two halves of the P-element are shown as diagonally striped boxes and the 31bp inverted repeats as arrowheads. Some of the restriction sites of the polylinker (PL), represented as a white box, are specified below. To obtain the transposon P[(w)A112], the -14.6 to -2.7 *Bgl* II fragment (insert 2 of Fig. 6.1) was cloned into the *Bam* HI site of the polylinker of *pW8*. The orientation of the insertion is shown below, restriction sites of the polylinker are depicted on a thin line, restriction sites of the insert on a solid line.

The *LB20* construct

Breakpoints which characterize the *LB20* complementation group were mapped to the genomic DNA area from position -10 to 6.5. Within this area only one transcription unit had been discovered, corresponding to the genomic *Hind* III fragment from 0 to 1.3. This transcript, of 600bp, seemed to be the likely transcription unit of the *LB20* locus.

However, it was also known that *LB20/Df(1)JA117* heterozygotes were lethal, yet the proximal breakpoint was located 6.5kb distal to the 600 bp transcript. This suggested the possibility that the sequences within these 6.5kb DNA might be necessary for the proper functioning of the *LB20* locus. It was therefore concluded that a transformation construct would have to be designed which included the 6.5kb DNA region plus the 600 bp transcript (Fig 6.1, 6.3).

The 11kb *Bam* HI fragment illustrated in Figure 6.1 satisfied these criteria and was also separated from the potential *A112* transcription unit. The plasmid pP[(w)LB20] was constructed by cloning the 11kb *Bam* HI fragment into the *Bam* HI site of the *pW8* vector (Fig.6.3).

6.3.2 Transformation

The constructs pP[(w)LB20] and pP[(w)A112] were identified prior to amplification by means of colony hybridization. Plasmid DNA, prepared from hybridizing colonies, was digested with several restriction endonucleases to establish that the correct restriction fragments had been cloned into the *pW8* vector.

Figure 6.3 Design of the *LB20* transformation construct

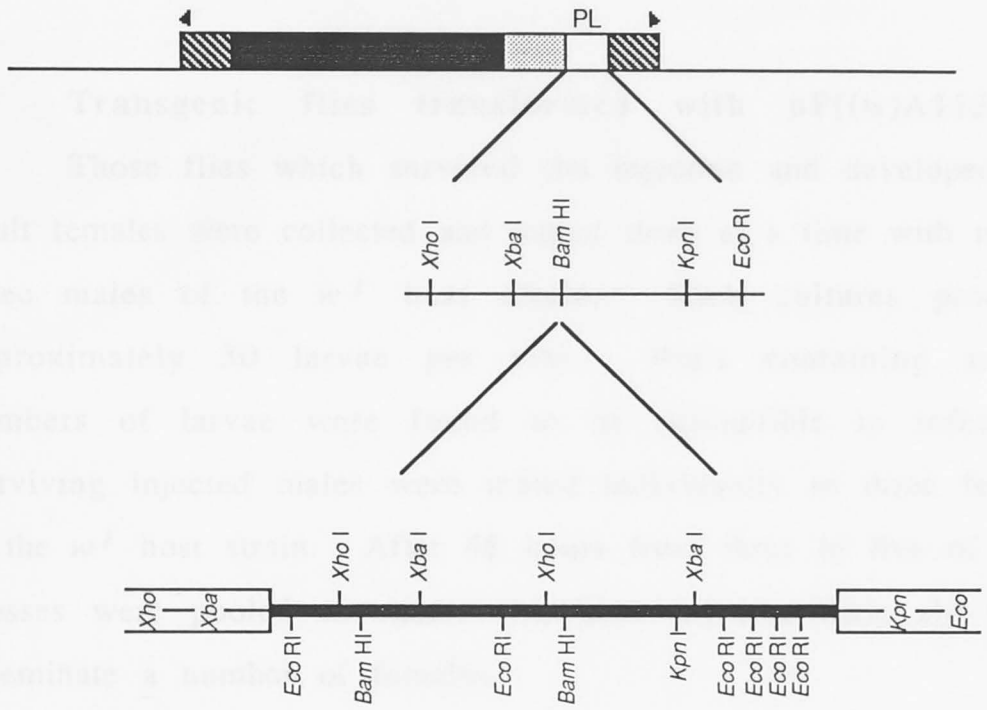


Fig. 6.3 Design of the transposon P[(w)LB20]. The *pW8* vector is illustrated at the top. The *white* gene is indicated as filled box and the *hsp 70* gene as dotted box. The two halves of the P-element are shown as diagonally striped boxes and the 31bp inverted repeats as arrowheads. Some of the restriction sites of the polylinker (PL), represented as a white box, are specified below. To obtain the transposon P[(w)LB20], the -8.6 to 2.4 *Bam* HI fragment (insert 1 of Fig. 6.1) was cloned into the *Bam* HI site of the polylinker of *pW8*. The orientation of the insertion is shown below, restriction sites of the polylinker are depicted on a thin line, restriction sites of the insert on a solid line.

The plasmids were coinjected with the P-factor helper plasmid pUC hs $\pi\Delta$ 2-3 (Mullins *et al.*, 1989) into embryos of a white-eyed fly strain homozygous for the mutation w^1. 1000 embryos were injected with pP[(w)LB20]. 30% of these developed into larvae and gave rise to 160 surviving adults. The equivalent number of embryos injected with pP[(w)A112] gave rise to 107 adults. Three transformant red-eyed flies were obtained in both cases, resulting in a transformation efficiency of 2% for pP[(w)LB20] and 3% for pP[(w)A112]. Although this transformation efficiency is

not high, it is not uncommon for transposons as large as 20kb (Spradling 1986).

Transgenic flies transformed with pP[(w)A112]

Those flies which survived the injection and developed into adult females were collected and mated three at a time with two to three males of the w^1 host strain. Such cultures produced approximately 30 larvae per vial. Vials containing smaller numbers of larvae were found to be susceptible to infections. Surviving injected males were mated individually to three females of the w^1 host strain. After 48 hours from three to five of these crosses were pooled to ensure that each injected male fly would inseminate a number of females.

Two red-eyed males, termed pink and black, and one red-eyed female, termed brown, were discovered in the subsequent G1 generation (Fig. 6.4), each originating from different vials.

Transgenic flies transformed with pP[(w)LB20]

Flies injected with pP[(w)LB20] were collected and mated in the same manner as those injected with pP[(w)A112]. Three transgenic males transformed with pP[(w)LB20] were discovered in the subsequent G_1 generation (Fig. 6.4). Two of those, termed 'blue' and 'green', are independent transformants since they originated from two separate vials. The third transformant, termed 'purple', was found in a vial holding the pool of flies which included the parents of the two transformants mentioned above. Therefore it

Figure 6.4: P element mediated transformation

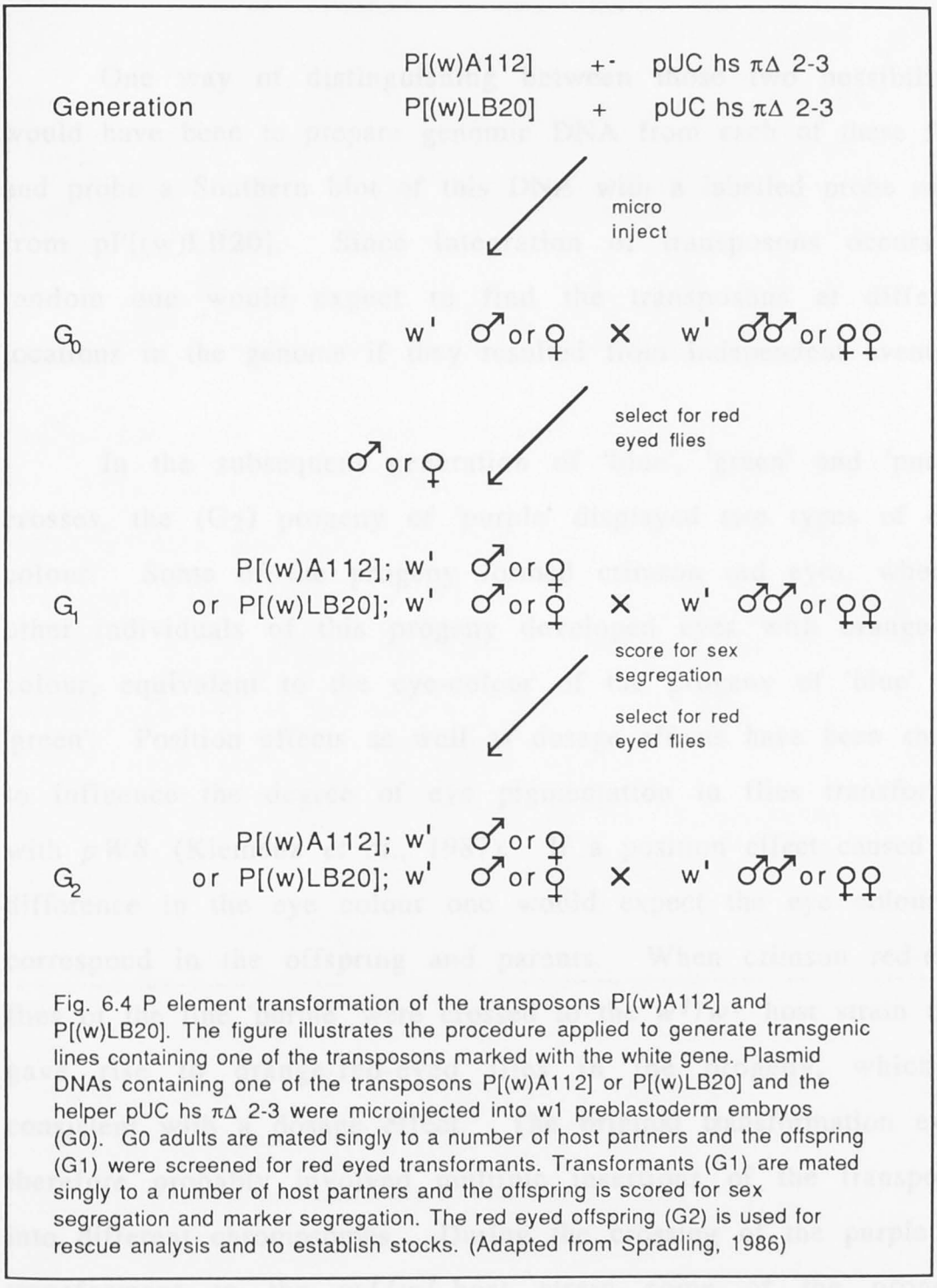


Fig. 6.4 P element transformation of the transposons P[(w)A112] and P[(w)LB20]. The figure illustrates the procedure applied to generate transgenic lines containing one of the transposons marked with the white gene. Plasmid DNAs containing one of the transposons P[(w)A112] or P[(w)LB20] and the helper pUC hs πΔ 2-3 were microinjected into w1 preblastoderm embryos (G0). G0 adults are mated singly to a number of host partners and the offspring (G1) were screened for red eyed transformants. Transformants (G1) are mated singly to a number of host partners and the offspring is scored for sex segregation and marker segregation. The red eyed offspring (G2) is used for rescue analysis and to establish stocks. (Adapted from Spradling, 1986)

might be possible that this transgenic fly is the offspring of one of the former two transformation events. However, because of the multiplicity of the germ line precursor cells, P-element DNA integrated more than once into the genome of a single pole cell can

give rise to two individuals with independent insertions in the progeny of that embryo.

One way of distinguishing between those two possibilities would have been to prepare genomic DNA from each of these flies and probe a Southern blot of this DNA with a labelled probe made from pP[(w)LB20]. Since integration of transposons occurs at random one would expect to find the transposons at different locations in the genome if they resulted from independent events.

In the subsequent generation of 'blue', 'green' and 'purple' crosses, the (G_2) progeny of 'purple' displayed two types of eye-colour. Some of the progeny formed crimson red eyes, whereas other individuals of this progeny developed eyes with orange-red colour, equivalent to the eye-colour of the progeny of 'blue' and 'green'. Position effects as well as dosage effects have been shown to influence the degree of eye pigmentation in flies transformed with *pW8* (Klemenz *et al.*, 1987). If a position effect caused the difference in the eye colour one would expect the eye colour to correspond in the offspring and parents. When crimson red-eyed flies of the line 'purple' were crossed to the w^1/w^1 host strain they gave rise to orange-red-eyed flies in the progeny, which is consistent with a dosage effect. The original transformation event therefore probably involved multiple insertions of the transposon into different chromosomes. During the crossing of the purple G_1 transformant to the w^1/w^1 host strain some of the progeny acquired both chromosomes with an insertion while some individuals received only one of those chromosomes. Individuals with one insertion, exhibiting orange eyes, developed only orange-eyed and white-eyed progeny when crossed to the w^1 host strain.

Assuming that individuals with two chromosomes bearing an insertion will exhibit crimson eyes one would expect in the progeny of such an individual out-crossed to the w^1 host strain, the combination of crimson red eyes, orange eyes and white eyes. However, most of the progeny of this cross proved to be inviable. Due to time constraints this phenomenon was not investigated further.

6.3.3 Analysis of transformants

In order to test whether the DNA introduced into these transformants was able to rescue the mutant phenotype it was important to establish that the insertion of the transposon did not occur on the X-chromosome. Each of the transgenic flies (G_1) was mated separately to individuals of the w^1 host strain. The offspring (G_2) were scored to establish whether the insertion of the transposon took place on one of the sex chromosomes or on an autosome. None of the crosses revealed linkage of the eye colour to sex. In all cases except one, the percentage of red-eyed and white-eyed males and females was found to be close to 25% (Table 6.1) and not significantly different to expectation, indicating an autosomal location for the transforming sequences. In the progeny from the pP[(w)LB20] plasmid there was an excess of red-eyed males and females (X^2_3 = 9.57, $p < 0.05$, >0.01). This result may be due to viability effects.

Table 6.1

		males				females				tot
		R		W		R		W		
plasmid	lines	No	%	No	%	No	%	No	%	
	brown	54	21.4	64	25.4	67	26.6	67	26.6	252
pP[(w)A112]	black	127	25.5	140	28.0	106	21.2	126	25.3	499
	pink	125	27.1	125	27.1	99	21.5	112	24.3	461
	purple	109	27.0	93	23.1	106	26.3	95	23.6	403
pP[(w)LB20]	blue	94	29.3	71	22.1	90	28.0	66	20.1	321*
	green	81	22.6	86	24.0	105	29.0	86	24.0	358

Table 6.1 Segregation of eye colour for the transformant lines. The first column indicates the plasmid used for transformation and the second lists the transgenic line tested. Each transformant (G1) was crossed several times to partners of the w^1 host strain and the offspring (G2) were scored. (The transformant line "brown" was tested with a G2 male). The number of males with red eyes (R) are given in column 3 and are also expressed as a percentage of the total number of scored flies in column 4 (tot). Similarly, the white-eyed (W) males are presented in numbers (column 5) and percentage (column 6), as well as the red-eyed (columns 7+8) and white-eyed females (columns 9+10). *Note significant departure from expectation.

The following experiment was designed to test whether the DNA sequences introduced into the transformants were able to complement the lethality of the *A112* and *LB20* mutant phenotypes. Available for testing were two alleles of the *A112* locus, $l(1)A112^{17\text{-}62}$ and $l(1)A112^{GF314}$, and three mutant alleles of the *LB20* locus, $l(1)LB20^{LB20}$, $l(1)LB20^{C27}$ and $l(1)LB20^{DA618}$. Two of those alleles, $l(1)A112^{GF314}$ and $l(1)LB20^{DA618}$, carry a second site lethal (George Miklos, personal communication) and were hence not suitable for the analysis.

Virgin females, of the remaining three lines, *l(1)A112$^{17-62}$ / Binsn*, *l(1)LB20^{LB20} / FM6* and *l(1)LB20^{C27} / FM6* were collected and mated to males of the transformant lines (Fig. 6.5).

Figure 6.5

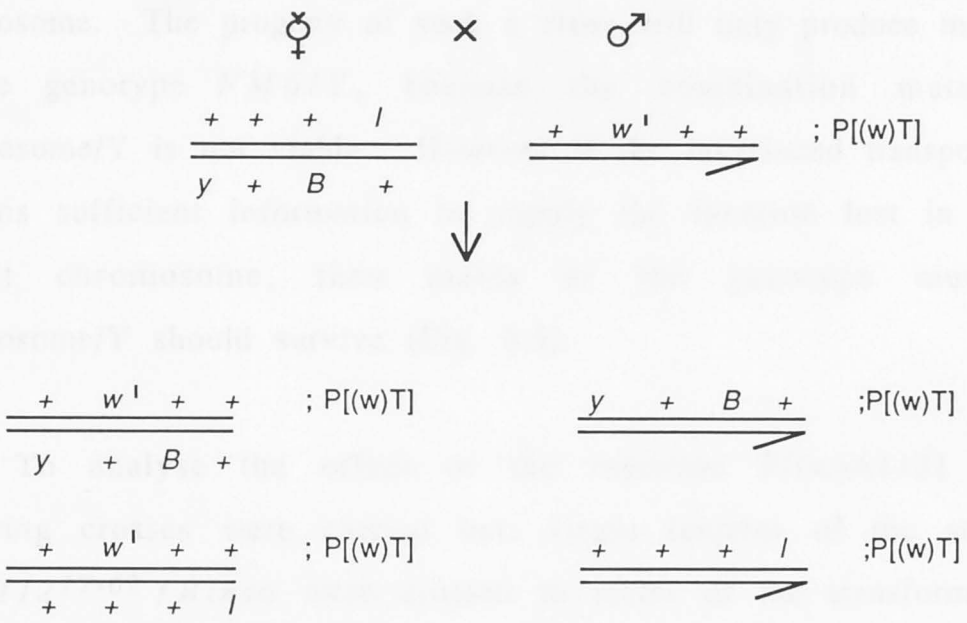


Figure 6.5 Crossing experiment of transformants carrying the transposon P[(w)T], to test for the ability to complement lethality. Virgins from a stock carrying an X-linked lethal (indicated as *l*) are crossed to males transformed on an autosome with the sequences P[(w)T]. Two types of X-chromosomes can pair with a Y-chromosome. One is marked with *yellow* and *Bar*, the other X-chromosome carries the lethal mutation. Males carrying the lethal mutation on their X-chromosome can only survive if they carry an autosome transformed with P[(w)T]. These males can be distinguished from the others because of their wild-type phenotype.

The experiment is based on the fact that the mutants studied are homozygous lethal. Because the alleles *l(1)LB20^{LB20}*, *l(1)LB20^{C27}* and *l(1)A112$^{17-62}$* are X-linked lethals, males carrying these alleles require a rearranged Y-chromosome such as *y^{+}Ymal^{+}*

or $y^{+}Ymal^{106}$ bearing sections of the X-chromosome (Schalet and Lefevre, 1976), to compensate for the lethality on the X-chromosome.

A cross between a virgin female of one of these mutant lines and a transformant male does not contain such a rearranged Y-chromosome. The progeny of such a cross will only produce males of the genotype *FM6/Y*, because the combination mutant-chromosome/Y is not viable. However if the introduced transposon contains sufficient information to supply the function lost in the mutant chromosome, then males of the genotype mutant chromosome/Y should survive (Fig. 6.5).

To analyse the effect of the insertion P[(w)A112] the following crosses were carried out: virgin females of the stock $l(1)A112^{17\text{-}62}$ / *Binsn* were crossed to males of the transformant lines 'black', 'brown' and 'pink', bearing the transposon P[(w)A112] (Table 6.2).

From these crosses I scored the male progeny for eye-colour. These data (Table 6.2) show that this transposon did complement the mutation at the *A112* locus. The offspring of these matings contained phenotypically normal males which carried the *A112* mutation on the X-chromosome.

Additionally, transformant males carrying the transposon P[(w)A112] were also crossed to virgin females bearing a lethal mutation at the *LB20* locus (Table 6.2). Phenotypically normal

Table 6.2

plasmid	line	mutant allele	males	No. crosses
		LB20	118	4
P[(w)A112]	'black'	*C27*	84	4
		17-62	81	4
		LB20	41	2
P[(w)A112]	'brown'	*C27*	4	1
		17-62	49	3
		LB20	70	4
P[(w)A112]	'pink'	*C27*	42	3
		17-62	52	3

Table. 6.2 Rescue experiment of the *A112* allele *17-62*. The transposon introduced is indicated in column 1. Column 2 lists the transgenic strains which were crossed to the different alleles of the *LB20* locus (*LB20*, *C27*) and the *A112* locus (*17-62*) cited in column 3. The number of phenotypically wild-type males which were found in the subsequent generation are listed in column 4. Each experiment was set up several times as indicated in column 5.

males were observed in all crosses of transformant lines with the *A112* allele ($l(1)A112^{17\text{-}62}$) as well as the *LB20* alleles ($l(1)LB20^{LB20}$ and $l(1)LB20^{C27}$). These surviving males are rescued by the transposon P[(w)A112] which complements the lethal mutation on their X-chromosomes. Hence the P[(w)A112] DNA covers both complementation groups, *LB20* as well as *A112*.

To analyse the effect of the insertion of P[(w)LB20] the following crosses were carried out (Table 6.3). Virgin females from the stocks $l(1)LB20^{LB20}/FM6$ and $l(1)LB20^{C27}/FM6$ were crossed to

males of the transformant lines 'purple', 'green' and 'blue', bearing the transposon P[(w)LB20].

Table 6.3

plasmid	line	mutant allele	wildtype males	No crosses
P[(w)LB20]	'green'	*LB20*	189	5
		C27	140	5
		17-62	0	4
P[(w)LB20]	'blue'	*LB20*	122	4
		C27	81	3
		17-62	0	4
P[(w)LB20]	'purple'	*LB20*	138	5
		C27	25	2
		17-62	0	4

Table. 6.3 Rescue experiment of the *LB20* alleles. The transposon introduced is indicated in column 1. Column 2 lists the transgenic strains which were crossed to the different alleles of the *LB20* locus (*LB20*, *C27*) and the *A112* locus (*17-62*) cited in column 3. The number of phenotypically wildtype males which were found in the subsequent generation are listed in column 4. Each experiment was set up several times as indicated in column 5.

Table 6.3 shows that this transposon did complement the mutation at the *LB20* locus. The offspring of these matings contained phenotypically normal males which carried the mutation on their X-chromosome. Flies with this genotype can only survive if the lethality on the mutant chromosomes $l(1)LB20^{LB20}$ (or $l(1)LB20^{C27}$) is complemented by the transposon. The cross of the *A112* allele $l(1)A112^{17\text{-}62}$ also mated to males from the three

transformed lines served as a control. No males developed from these crosses, indicating that the transposon P[(w)LB20] does not rescue the lethality of the *A112* locus.

6.4 Discussion

In this chapter P-element mediated transformation was used to analyse whether the molecular areas provisionally assigned to the loci *LB20* and *A112* (chapter 2) are sufficient to cover the biological function of these complementation groups. Several transgenic lines were obtained from w^1 flies transformed with either of two transposons, P[(w)A112] or P[(w)LB20]. These transposons were designed such that the inserts of genomic DNA overlap by 6kb and hence define three molecular areas, the overlap of 6kb as well as the proximal and distal regions. The results show that each of these molecular areas contains a different gene.

The *A112* gene was reintegrated into the *Drosophila* genome via P-element mediated germline transformation. Three independent transformants were obtained each with the transposon P[(w)A112] located on an autosome (Table 6.1). In all lines derived from these transformants the introduced DNA complemented $l(1)A112^{17\text{-}62}$ (a lethal allele of the *A112* gene) as demonstrated by the survival of adult males carrying this mutation on their X-chromosomes (Table 6.2). These rescued flies were phenotypically normal, appeared healthy and were fertile. Hence the 12kb genomic DNA fragment of this transposon P[(w)A112], from the genomic map at position -14.6 to -2.7 (Fig. 6.1), is sufficient to complement completely the lethal mutation of the *A112* gene.

Consistent with these data are the analyses previously described in chapters 2 and 3. A 2.2kb transcription unit was

observed proximal to position -8.6 and this transcription unit is interrupted by the proximal breakpoint of *Df(1)HM44* as well as the distal breakpoint of *Df(1)JA117*. These two deficiencies were used to pin-point the *A112* gene on the molecular map. This 2.2kb transcript is fully contained within the transposon P[(w)A112], which rescues the *A112* allele, but not within P[(w)LB20], which does not rescue the *A112* allele. Therefore the conclusion is that the transcription unit corresponding to the *A112* gene is the 2.2kb transcript.

A second transposon, P[(w)LB20], was also injected into embryos. This transposon gave rise to three independent transformants each leading to an identical result, namely they rescued the two alleles tested for *LB20* but not the allele tested for *A112* (Table 6.3). From these results it can be deduced that the limits of the *LB20* gene are contained within the DNA insert which was introduced into the flies. The only transcript detected initially was a 600nt transcript corresponding to the genomic 1.3kb *Hind*III fragment at position 0 to 1.3. This 600nt transcript was therefore a possible candidate to be the transcription unit of the *LB20* locus.

However, the P[(w)A112] transposon, when introduced into the germline of three flies not only rescued the *A112* allele but also both *LB20* alleles tested (Table 6.2). However, this transposon does not include the genomic DNA corresponding to the 600nt transcription unit, clearly indicating that this transcript cannot be from the *LB20* gene.

The design of the transposons was based on the information of transcriptional activity (chapter 3) sequence data (chapter 4 and 5) and the mapped position of chromosomal deficiencies (chapter

2). Although only the 600nt transcript had been detected the breakpoint of *Df(1)JA117* was evidence that the sequences distal to this 600nt transcript are necessary for the *LB20* gene function.

As the genomic DNA contained within those two transposons overlaps by 6kb, the most likely explanation for the complementation of the *LB20* alleles is therefore that the transcription unit corresponding to the *LB20* gene resides within this 6kb overlap and was probably missed in the transcriptional analysis because of its low abundance. This could be investigated using polyadenylated RNA, although it is possible that the transcript is at a very low level and is undetectable using these techniques.

Chapter 7

General Discussion

7. General Discussion

This thesis is concerned with genetic and molecular investigations of two recessive lethal genes which map to the adjacent complementation groups *A112* and *LB20*. In the preceding five chapters, data have been presented which focus on the location, structure and function of the genes in the *A112/LB20* region. These data include mapping the genes at the genomic level using chromosomal deficiencies, as well as at the transcriptional level by analysing the transcripts produced in the region. The subsequent isolation of cDNAs with homology to the genomic region allowed a detailed analysis of the structure of the two genes. Examination of the nucleotide sequence specified the characteristics of the "collagen-like" gene and identified the RNA helicase function of the *A112* gene. Finally, germline transformation using two constructs whose design was based on the previously acquired information, confirmed the mapping of the *A112* gene further defined the location of the *LB20* gene and led to the discovery that there are probably three genes in the region.

The results of my experiments have revealed more about the *A112* gene than they have about the *LB20* gene. The *A112* gene has been mapped to the genomic region from -13 to -6.8. Within that region only a single transcript was detected. This transcript (of 2.2kb) is expressed in embryonic, larval, pupal and adult tissue.

A cDNA homologous to this transcript was isolated and sequenced. Comparison of the predicted amino acid sequence to those of known proteins showed that the *A112* gene clearly belongs to a recently defined family of RNA helicases. The A112 protein sequence contains the specific motifs which characterise this family (Gorbalenya *et al.*, 1988; Hodgman, 1988) including the DEAD box which is a motif important in ATP-binding and in this version unique to RNA helicases (Linder et al., 1989; Chang *et al.*, 1990).

The homology of the cDNA to the genomic *A112* region was confirmed by hybridisation of the labelled cDNA to genomic DNA and also by comparing the sequences of the cDNA to some genomic sequences, which led to the discovery of two small introns at the 5'end. In some RNA helicases, like *vasa* for example, seven exons have been reported separated by 5 introns, four of which are only 53bp to 70bp long (Hay *et al.*, 1988; Lasko and Ashburner, 1988). It cannot be excluded that there are more introns present in the *A112* gene, since the genomic sequences are not complete. Also, because no stop codon was found in the coding frame upstream of the first methionine, it cannot be discounted that the cDNA might not be quite full length. However, the cDNA is only 250 nucleotides shorter than the corresponding transcript measured on Northern blots, and it is obvious that the missing region has to be very small.

The successful transformation and subsequent rescue experiments described in chapter 6 provided proof that the *A112* gene is located within the introduced 12 kb fragment (from position -14.6 to -2.7). This was shown with three independent transgenic lines. This analysis provides a very good basis from

which to investigate the regulation of the *A112* gene. Very little is known about the regulation of RNA helicases, since this family is so diverse. Hence it would be extremely interesting to elucidate the control mechanisms of the *A112* gene. For example, a reporter gene like chloramphenicol acetyl transferase (CAT) or the *lac Z* gene, fused to various 5' upstream fragments from the *A112* gene, could be used to determine which sequences are involved in its regulation. Deletion analyses have been successfully used to define 5' cis acting regulatory sequences in a wide variety of genes, for example the yolk protein gene (*yp1*) (Garabedian *et al.*, 1986), the major Larval Serum Protein (*Lsp-1*) (Delaney, 1987), the gap gene *Krüppel* (*Kr*) (Hoch *et al.*, 1990) or the *twist* gene (*twi*) (Thisse *et al.*, 1991). In this form of analysis the upstream sequences are successively reduced, fused to reporter genes and the effect on expression is analysed *in vivo* using P-element transformation of the constructs. The CAT construct might be particularly useful in attempting to identify the regulatory cis-acting sequences because of the high sensitivity of this assay.

Although RNA helicases have in common the catalytic activity of unwinding double stranded RNAs, their biological functions are quite different. For instance, mammalian eIF-4A and its counterparts in yeast, TIF1 and TIF2, are essential in translation initiation (Hay *et al.*, 1988; Nielsen and Trachsel, 1988; Linder and Slonimski, 1989), whereas p68 is thought to be involved in the regulation of cell growth and division (Ford, *et al.*, 1988). Other RNA helicases play a role in mRNA splicing (Seraphin, *et al.*, 1989) and ribosome assembly (Nishi, *et al.*, 1988). It has been suggested that the sequence conservation of the RNA helicase family, spanning enzymes from eubacteria to eukaryotes, is an indication of its early evolution (Chang *et al.*, 1990) and that the RNA helicase

activity which played an important part in self-catalytic replication (Gilbert, 1986) then developed into many specialised roles that can be observed in present day organisms. Thus it is difficult to predict from the sequence similarity the biological function of the *A112* gene, other than that it is concerned with the unwinding of double stranded RNAs. It has been suggested that the specific biological role of the individual proteins is encoded outside the RNA helicase region, in the C-terminal and or N-terminal parts of the proteins (Chang *et al.*, 1990). This idea is supported by the lack of sequence similarity between different proteins outside the RNA helicase region. A similar characteristic exists in the family of Rel proteins, whose members also share a 300 amino acid domain of highly related residues but in addition have individual protein termini with little similarity (Gilmore, 1991).

The developmental expression of the *A112* RNA shows a slightly increased expression in embryonic and adult tissue, but basically the gene is expressed in all stages tested, which agrees well with the proposed function of a RNA helicase. It can be assumed that the biological function of the *A112* complementation group is essential for viability since every mutant allele discovered at this locus is homozygous lethal and germline clone analysis revealed that mutations at this locus are germ cell lethal (Perrimon *et al.*, 1989). This is not necessarily a contradiction to the proposed function of a RNA helicase. Similar to the example in yeast (Linder *et al.*, 1989; Chang *et al.*, 1990), where the *TIF1* and *TIF2* genes, also members of the RNA helicase family, affect cell viability and mitochondrial functions if both inactivated, one can expect

numerous RNA helicases in *Drosophila*, each with a specific function which might or might not be connected to viability.

Apart from further genetic and molecular characterisations of *A112* it would be of interest to isolate the encoded protein and study its biochemical function. For this it would be necessary to raise antibodies against a fusion protein and then purify the protein (Snyder, 1989). It would be then possible to test whether the purified protein has an RNA unwinding activity and whether this activity is ATP dependent, as was shown for p68 (Hirling, *et al.*, 1989), SrmB (Nishi et al., 1988) and Prp 16 (Schwer and Guthrie, 1991).

My study has not provided such a complete description of the *LB20* gene. The cytological borders of the *LB20* locus are on the X-chromosome in 19F, proximal to the proximal breakpoint of *Df(1)HM44* (Miklos *et al.*, 1986). The molecular borders of this locus were mapped proximal to position -10 and distal to position 6.5 on the molecular map. Within that region two transcripts were discovered; a 0.6kb transcript corresponds to position 0 to 1.3 and, within the genomic fragment proximal to that, a second transcript of 0.4kb was located. The interpretation as to which transcript might be the one corresponding to the *LB20* locus was influenced by the position of the proximal breakpoint of *Df(1)JA117*, which mapped to position -6.4 on the molecular map. Because this deficiency is lethal in heterozygotes with mutant alleles at the *A112* and *LB20* loci, but viable with alleles carrying lethal mutations proximal to *LB20* (Lefevre, 1981), it was clear that the DNA proximal to both transcripts was important for the proper

LB20 function and subsequently led to the decision that the 0.6kb was a more likely candidate for the *LB20* transcription unit. The transformation experiments showed however, that this transcript was not relevant to *LB20* function. The information on breakpoints, transcripts and transformation constructs is summarised in figure 7.1.

After the transformation experiments had been done it was apparent that, since the Northern analyses were carried out with total RNA, it was possible that if a transcript was relatively rare it may well have been missed, as for example, transcripts associated with a number of the eye pigment genes in *D. melanogaster* are known to be produced in very low copy number like the *white* (O'Hare, *et al.*, 1983) or *scarlet* transcripts (Tearle, 1987). In an effort to increase the sensitivity of the experiment I isolated poly(A^+)RNA using oligo (dT) columns. A Northern blot using this poly(A^+)RNA was carried out to screen the region once more for transcripts (Fig. 7.2). Because both transformation constructs contain sufficient DNA information to completely rescue the defect of lethal mutations at the *LB20* locus, the relevant DNA has to be within the overlap of those two constructs. This reduces the possible DNA region to the 5.9 kb (Fig. 7.1) which extends from position -8.6, the distal end of the *LB20* construct, to position -2.7, the proximal end of the *A112* construct. Three probes were isolated from this region, and one of these, the 2.2kb *Bam* HI/ *Kpn* I fragment from position -8.6 to -6.4 was found to hybridise indeed to a band faintly visible at 3kb in the lane containing poly(A^+)RNA from adult tissue (Fig. 7.2). The other band visible at 2.2kb corresponds to the *A112* transcript. The nucleotide sequence of the 2.2kb transcript shows a *Bam* H1 restriction site at

nucleotide 235 and a *Sal* I restriction site at nucleotide 178, corresponding to the genomic *Bam* H1 restriction site at position -8.6 and the *Sal* I restriction site at position -8.5 on the genomic map. Therefore, the new 3kb transcript has to be located proximal to the *Sal* I restriction site at -8.5 and distal to the *Bgl* II restriction site at -2.7.

Figure 7.1

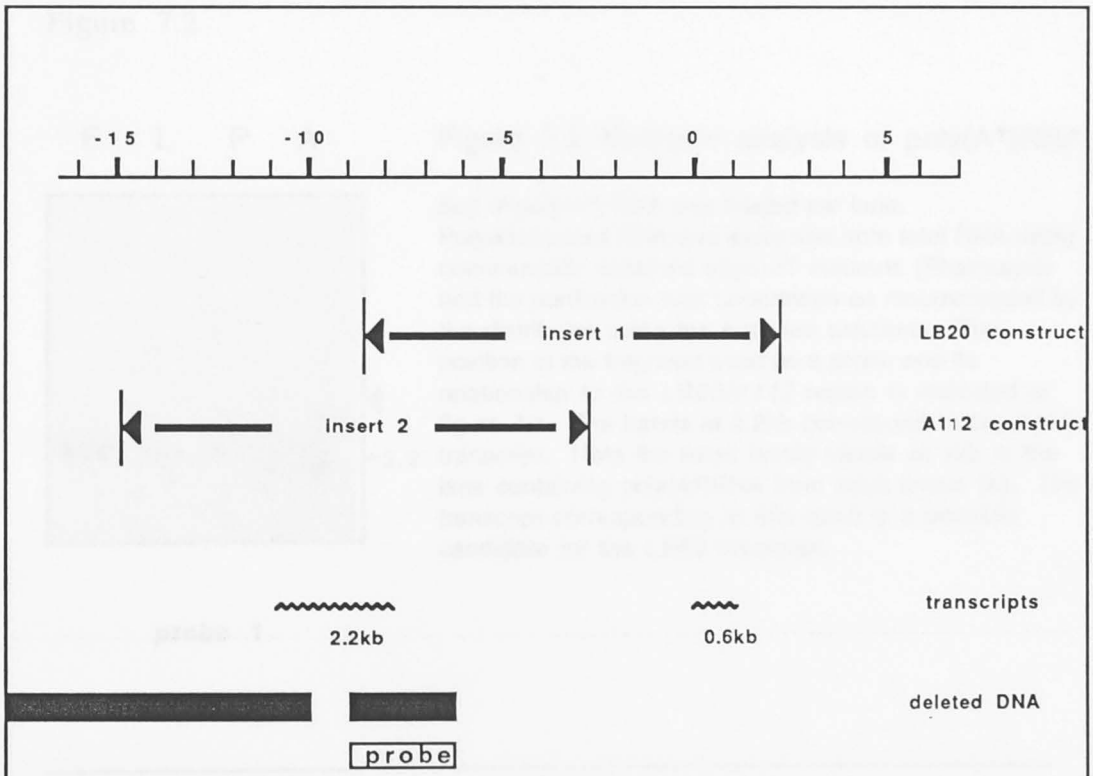


Fig.7.1 The *LB20/A112* region.
The relationships between genomic DNA, transcripts, transposons and relevant deficiencies is illustrated.
The co-ordinate system above corresponds to Fig. 2.1. Arrows indicate the limits of the genomic fragments used to construct the transposons. Insert 1 is contained within the transposon P[(w)LB20], insert 2 in transposon P[(w)A112]. The two transcripts of 2.2kb and 0.6kb are indicated as wavy lines. The black squares represent deleted DNA in two deficiencies: the deleted DNA at the proximal breakpoint of *Df(1)HM44* is indicated to the left, the deleted DNA in *Df(1)JA117* is indicated to the right. The white square illustrates the position of the probe discussed in figure 7.2.

Two conclusions can be drawn from these preliminary data. First, it seems likely that this 3kb transcript is the *LB20* transcript. This could be confirmed by the isolation of a cDNA corresponding to this transcript. A second conclusion can be drawn with respect to the location of the *LB20* transcript. The outer limits of the *LB20* gene can now be defined even more precisely than by the borders

---

Figure 7.2

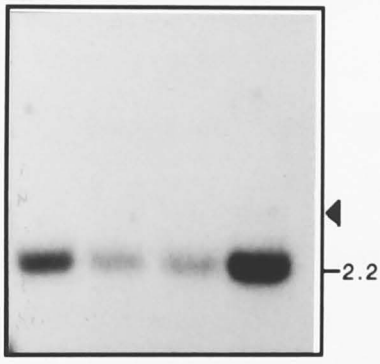


Figure 7.2 Northern analysis of poly(A$^+$)RNA

5µg of poly(A$^+$) RNA was loaded per lane. Polyadenylated RNA was extracted from total RNA using commercially obtained oligo dT columns (Pharmacia) and the purification was undertaken as recommended by the distributer using the supplied solutions. The position of the fragment used as a probe and its relationship to the *LB20/A112* region is indicated in figure 7.1. The bands at 2.2kb correspond to the *A112* transcript. Note the band faintly visible at 3kb in the lane containing poly(A$^+$)RNA from adult tissue (A). The transcript corresponding to this band is a possible candidate for the *LB20* transcript.

---

of the transformation constructs. The *A112* transcript impinges on this DNA region (the genomic *Sal* I restriction site at -8.5 corresponds to the *Sal* I restriciton site 178bp downstream of the 5'end of the cDNA). In addition to the known *A112* sequence it is likely that there is a region 5' that contains the 40 to 110 residues for the TATA box sequences and upstream control elements (Hultmark, 1986; Buratowski *et al.*, 1989). In yeast, upstream

sequences are found between 20 and 500 bp upstream of the TATA box (Struhl, 1987) and in higher eukaryotes they can usually be found within 500bp to the TATA box (Serfling *et al.*, 1985). Thus, the *LB20* gene is possibly located upstream to position -8. Thus, using P-element transformation it was possible to define the borders of the *LB20* gene to approximately 5kb. A strategy to map the region in even more detail would be to design new transformation constructs, smaller than the *A112*- and *LB20*- constructs used, and to repeat the germline transformations.

However, the more important experiment would be to screen a cDNA library, made from adult tissue, with the 2.2kb *Bam* HI/ *Kpn* I fragment from position -8.6 to -6.4 as a probe, to isolate a cDNA corresponding to this *LB20* transcript. The hybridisation of this cDNA to the genome would determine the precise borders of the *LB20* gene. Even more importantly, the cDNA could be sequenced. Because it is now known that alleles of the *LB20* locus are mitotic mutants (Mauritio Gatti, personal communication 1989; Masatoshi Yamamoto, personal communication 1989) it is possible that comparing the sequence to known mitotic genes might reveal the precise function of the *LB20* gene.

The transformation experiments described in chapter 6 clearly showed that the transcript corresponding to position 0 to 1.3, that I initially identified as the "collagen-like" gene and sequenced, cannot be from the *LB20* gene since sequences

corresponding to this transcript are not required for the *LB20* function. This transcript therefore belongs to the proximally adjacent locus.

The next known complementation group proximally adjacent to *LB20* is *tumorous head tuh-1* (Woolf, 1968; Pyati, 1976). The breakpoints of two deficiencies, *Df(1)JC4* and *Df(1)Q539*, which border the *tuh-1* locus (Pyati, 1976), show that the "collagen-like" gene cannot be part of this locus. Flies of the genotype *LB20/Df(1)JC4* are phenotypically wild-type (Lefevre, 1981), but *tuh-1/Df(1)JC4* (Pyati, 1976) flies express the phenotype of *tuh-1*. From this information one would expect that transcripts of the *tuh-1* locus are deleted in *Df(1)JC4*. The molecular breakpoint of this deficiency was mapped approximately 40kb proximal to the *pCog* transcript.

A similar result is obtained with *Df(1)Q539* . Heterozygous flies of the genotype *LB20/Df(1)Q539* (Schalet and Lefevre, 1976) are lethal but *tuh-1/Df(1)Q539* flies are phenotypically normal (Pyati, 1976), hence DNA from the *LB20* gene is deleted in this deficiency, but DNA from the *tuh-1* locus is not. The breakpoint of this deficiency was shown to be proximal to the *pCog* transcript. Because the deficiency mapping shows that the *pCog* transcript is deleted in this deficiency it follows that the *pCog* transcript is not part of the *tuh-1* complementation group. Thus, the conclusion is that the "collagen-like" gene belongs to a new locus, which exists between the loci *LB20* and *tuh-1*.

The transcriptional analysis shows that the "collagen-like" gene is strongly expressed in adult tissue (chapter 3), less strongly in larval tissue and comparatively weakly in the embryonic and

pupal tissues. This developmental profile is also found with poly(A^+)RNA (data not shown). The sequence analysis also showed that transcriptional initiation signals are present. A TATA box was found in good agreement with the consensus sequences reported in the literature (Corden *et al.*, 1980; Manley, 1988). The start consensus site is also in agreement with the consensus sequence calculated for the start of transcription in *Drosophila* (Hultmark *et al.*, 1986). This is further evidence that the "collagen-like" gene is not a pseudogene but is expressed and transcribed. Processed pseudogenes, which have been reported mainly in mammals (Wagner, 1986), lack the promoter sequences upstream to the cap site and are flanked by direct repeats (Jeffreys and Harris, 1984; Vanin, 1984). It was also shown from the amino acid analysis that the protein translated from the "collagen-like" gene had a signal peptide at the 5' end (Perlman and Halvorson, 1983), which would only be expected in a functional gene.

The sequence of the "collagen-like" gene revealed that it contained a repeat with the basic motif "GGX GGX CCX". The amino acid translation of this repeat results in the basic amino acid sequence glycine-glycine-proline. A comparison of the predicted amino acid sequence encoded by the "collagen-like" gene to known proteins revealed initially a similarity to collagens, which was based on this motif. However, this was only a superficial similarity caused by the glycine-glycine-proline repeat of the "collagen-like" gene which resembles the central characteristics of the collagen protein family but does not coincide with them. These characteristics are that within the triple helical domain the amino acid glycine is always the third residue, and additionally within the amino acid motif glycine-X-Y, X and Y are non-identical amino

acids but X is often a proline (Ramachandran, 1967; Fessler and Fessler, 1978; Bornstein and Sage, 1980; Le Parco *et al.*, 1986).

Although this sequence similarity is not significant at the functional level, it was most probably the basis for the initial hybridisation of the chicken collagen probe to the *Drosophila* clone *DCg2* (Natzle *et al.*, 1982), based on the triple helical domain. It is interesting in this context that attempts to hybridise this *Drosophila melanogaster* clone *DCg2* (Natzle *et al.*, 1982) to the salivary gland chromosomes of *Drosophila virilis* were not successful (Whiting, *et al.*, 1989).

The amino acid repeat glycine-glycine-proline, although not perfect, contains 97.5 % of the two amino acids glycine and proline, which permits the triple helical structure in collagens (Fleischmajer *et al.*, 1985). Because of the high concentration of these amino acids the prediction is that the "collagen-like" gene could have an helical structure comparable to the triple helical structure of the collagen α-chains. It would be interesting in this context to investigate whether the "collagen-like" protein also forms trimers, like those of the collagen α-chains, which form the rod-like molecule. If an antibody against a fusion protein of the "collagen-like" gene were raised, it would be possible to isolate the protein and investigate its biochemical properties.

The rod-like collagen molecule is assembled from three collagen α-chains which have been secreted into the extracellular matrix (Bornstein and Sage, 1980; Olsen, 1981). It is intriguing that the "collagen-like" protein might also be secreted into the extra-cellular matrix. This was suggested by the fact that the first eighteen amino acids have the properties of a signal peptide

(Perlman and Halvorson et al., 1983), which is used as a transport signal to pass through the endoplasmic reticulum and from there via the Golgi apparatus, into the extra-cellular matrix (Garoff, 1985; von Heijne, 1985; Wiedman *et al.*, 1987). Thus it would be very interesting to analyse the inter-cellular distribution of the "collagen-like" protein. Immunological labelling of tissue sections could elucidate the location of the protein and where and in what stages it occurs. *In situ* hybridisation on *Drosophila* tissue sections probed with fragments of *DCg1*, the other collagen clone isolated by Natzle *et al.* (1982), identified a specific accumulation of this type of collagen in cells referred to as haemocytes (Knibiehler *et al.*, 1987; Mirre *et al.*, 1988).

One question that remains is the nature of the lesions in the recessive lethal genes *A112* and *LB20*. Both sets of lethal alleles were produced by X-ray and hence it may be that each of these lethals has deletions. It would be interesting to determine the nature of the mutations. Such studies might provide further information as to which part of the proteins is essential for their function.

The results presented achieved a detailed description of the *A112/LB20* region on the *Drosophila* X-chromosome. Important steps have been made towards gaining an insight into the biological functions of the two lethal genes encoded in this area. The loci *A112* and *LB20* are in close proximity and it is now understandable why it has been difficult to find deficiencies which separate these loci. The evidence that there may be a transcription unit between *LB20* and *tuh-1* is not surprising given that there is 40kb of DNA between these loci. These data support the

observation of Lefevre and Watkins (1986) who have argued that the number of lethal complementation groups will underestimate the gene density of the region.

In conclusion, the experimental results described in this thesis underline the scientific value of combining genetic and molecular approaches to reveal the biological nature of mutant genes hitherto known only from their severe effects on viability.

Bibliography

Bibliography

Adams, I. (1978) Invertebrate Collagens, *Science* **202**:591-598

Akam, M. (1989) The molecular basis to metameric pattern in the *Drosophila* embryo, *Development* **101**:1-22

Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridisation with DNA probes, *Proc. Natl. Acad. Sci.* **74**:5350-5354

Auerbach, C., Robson, J.M. and Carr, J.G. (1947) The chemical production of mutations, *Science* **105**:243-247

Baer, R., Bankier, A.T., Biggin, M.D., Deininger, P.L., Farrell, P.J., Gibson, T.J., Hatfull, G., Hudson, G.S., Satchwell, S.C., Seguin, C., Tuffnell, P.S. and Barrell, B.G. (1984) DNA sequence and expression of the B95-8 Epstein-Barr virus genome, *Nature* **310**:207-211

Bankier, A. T. and Barrell, B. G. (1984) *Techniques in Nucleic Acids Biochemistry* (ed. Flavell) B508, 1 Elsevier Scientific Publishers, Ireland

Bentz, H., Morris, N.P., Murray, L.W., Sakai, L.Y., Hollister, D.W. and Burgeson, R.E. (1983) Isolation and partial characterization of a new human collagen with an extended triple-helical structural domain, *Proc. Natl. Sci.* **80**:3168-3172

Bender, W., Spierer, P. and Hogness, D.S. (1983) Chromosomal walking and jumping to isolate DNA from the *bithorax* complex and the *Ace* and *rosy* loci in *Drosophila melanogaster*, *J. Mol. Biol.* **168**:17-33

Birnboim, H.C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA, *Nucl. Acids Res.* **7**:1513-1523

Birnstiel, M., Busslinger, M. and Strub, A. (1985) Transcription termination and 3' processing: the end is in site, *Cell* **41**:349-359

Bornstein, P. and Sage, H. (1980) Structurally distinct collagen types, *Ann. Rev. Biochem.* **49**:957-1003

Bossy, B., Hall, L.M.C. and Spierer, P. (1984) Genetic activity along 315kb of *Drosophila* chromosome, *EMBO J.* **3**:2537-2541.

Bridges, C.B. (1935) Salivary chromosome maps, *J. Hered.* **26**:60-64

Bridges, C.B. (1938) A revised map of the salivary gland X-chromosome of *Drosophila melanogaster*, *J. Hered.* **29**:81-86

Brutlag, D.L., Clauton, J., Freidland, P. and Kedes, L.H. (1982) SEQ: a nucleotide sequence and analysis recombination system, *Nucl. Acids Res.* **10**:279-294

Buerglin, T.R. (1991) The TEA domain: a novel, highly conserved DNA-binding motif, *Cell* **66**:11-12

Buratowsky, S., Hahn, S., Guarente, L. and Sharp, A.P. (1989) Five intermediate complexes in transcription initiation by RNA polymerase II, *Cell* **56**:549-561

Burgess, S., Couto, J.R. and Guthrie, C. (1990) A putative ATP binding protein influences the fidelity of branchpoint recognition in yeast splicing, *Cell* **60**:705-717

Cavener, D.R., Otteson, D.C. and Kaufman, T.C. (1986) A rehabilitation of the genetic map of the 84B-D region in *Drosophila melanogaster*, *Genetics* **114**:111-123

Chang, C., Kokntis, J. and Liao, S. (1988) Molecular cloning of human and rat complementary DNA encoding androgen receptors, *Science* **240**:324-326

Chang, T-H., Arenas, J. and Abelson, J. (1990) Identification of five putative yeast RNA helicase genes, *Proc. Natl. Acad. Sci. USA* **87**:1571-1575

Chen, J.H. and Lin, R.J. (1990) The yeast PRP2 protein, a putative RNA-dependent ATPase, shares extensive sequence homology with two other pre-mRNA splicing factors, *Nucl. Acids Res.* **18**:6447

Chovnick, A., Finnerty, V., Schalet, A. and Duck, P. (1969) Studies on the genetic organization on higher organisms: I. Analysis of a complex gene in *Drosophila melanogaster*, *Genetics* **62**:145-160

Chovnick, A., Gelbart, W. and McCarron, M. (1977) Organization at the *Rosy* locus in *Drosophila melanogaster*, *Cell* **11**:1-10

Company, M., Arenas, J. and Abelson, J. (1991) Requirement of the RNA helicase-like protein PRP22 for release of messenger RNA from spliceosomes, *Nature* **349**:487-493

Corden, J., Waslyk, B., Buckwalder, A., Sassone-Corsi, A., Kedinger, P. and Chambon, P. (1980) Promoter sequences of eukaryotic protein coding regions, *Science* **209**:1406-1414

Dalbadie-McFarland, G. and Abelson, J. (1990) PRP5: A helicase-like protein required for mRNA splicing in yeast, *Proc. Natl. Acad. Sci. USA* **87**:4236-4240

Delaney, S. (1987) Cis-acting regulatory elements of the *larval serum protein-1* genes of *Drosophila*, PhD thesis, Imperial College of Science and Technology, London

Dente, L., Caesarini, G. and Cortese, R. (1983) *pEMBL*: a new family of single stranded plasmids, *Nucl. Acids Res.* **11**:1645-1655

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX, *Nucl. Acids Res.* **12**:387-395

Digan, M.F., Haynes, S.R., Mozer, B.A., Dawid, J.B., Forquignon, F. and Gans, M. (1986) Genetic and molecular analysis of *fs(1)h*, a maternal effect homeotic gene in *Drosophila*, *Dev. Biol.* **114**:161-169

Dorer, D.R., Christensen, A.C. and Johnson, D.H. (1990) A novel RNA helicase gene tightly linked to the *Triplo-lethal* locus of *Drosophila*, *Nucl. Acids Res.* **18**:5489-5494

Duboule, D. and Dolle, P. (1989) The structural and functional organization of the murine *Hox* gene family resembles that of *Drosophila* homeotic genes, *EMBO J.* **8**:1498-1505

Dutton, F.L. Jr. and Chovnick, A. (1991) The 1(3)*S12* locus of *Drosophila melanogaster*: Heterochromatic position effects and stage-specific misexpression of the gene in P element transposons, *Genetics* **228**:103-118

Edgar, B.A. and O'Farrell, P. (1989) Genetic control of cell division patterns in the *Drosophila* embryo, *Cell* **57**:177-187

Eeken, J.C.J., Sobels, F.H., Hyland, V. and Schalet, A.P., (1985) Distribution of MR-induced sex-linked recessive lethal mutations in *Drosophila melanogaster*, *Mutat. Res.* **150**:261-275

Engels, W.R. (1989) P elements in *Drosophila melanogaster*, In *Mobile DNA* (eds. Berg and Howe) pp 437-484, American Society for Microbiology

Feinberg, A.P. and Vogelstein, B. (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high activity, *Anal. Biochem.* **132**:6-13

Fessler, J.H. and Fessler, L.I. (1978) Biosynthesis of procollagen, *Ann. Rev. Biochem.* **47**:129-162

Fietzek, P.P., Allmann, H., Rautenberg, J., Henkel, W., Wachter, E. and Kuehn, K. (1979) Amino acid sequence of a bovine collagen alpha 1(III) chain, *Hoppe-Seyler's Z. Physiol. Chem.* **360**:809-820

Fischbach, K.F. and Heisenberg, M. (1981) Structural brain mutant of *Drosophila melanogaster* with reduced cell number in the medulla cortex and with normal optomotor response, *Proc. Natl. Acad. Sci. USA* **78**:1105-1109

Fleischmajer, R., Olsen, B.R., Kuhn, K. (eds) (1985) *Biology, Chemistry and Pathology of Collagen*. Ann. N.Y. Acad Aci, Vol 460

Ford, M.J., Anton, I.A. and Lane, D.P. (1988) Nuclear protein with sequence homology to translation initiation factor eIF-4A, *Nature* **332**:736-740

French, B.T., Lee, W.H. and Maul, G.G. (1985) Nucleotide sequence of a cDNA clone for mouse pro-alpha1(I) collagen protein, *Gene* **39**:311-312

Frohman, M.A., Dush, M.A. and Martin, G.R. (1988) Rapid production of full-length cDNA from rare transcripts: Amplification using a single gene-specific oligonucleotide primer, *Proc. Natl. Acad. Sci.* **85**:8998-9002

Garabedian, M.J., Shephard, B.M. and Wensink, P.C. (1986) A tissue specific transcription enhancer for the *Drosophila* yolk protein-1 gene, *Cell* **45**:859-867

Garoff, H. (1985) Using recombinant DNA techniques to study protein targeting in the eukaryotic cell. *Annu. Rev. Cell Biol.* **1**:403-445.

Gatti, M. and Baker, B.S. (1989) Genes controlling essential cell cycle functions, *Genes Dev.* **3**:438-453

Gaunt, S.J., Sharpe, P.T. and Duboule, D. (1988) Spatially restricted domains of homeogene transcripts in mouse embryos: relation to a segment body plan, *Development* **104**:(Suppl.): 169-175

Gausz, J., Bencze, G., Gyurkovics, H., Ashburner, M., Ish-Horowicz, D. and Holden, J.J. (1979) Genetic characterization of the 87C region of the third chromosome in *Drosophila melanogaster*, *Genetics* **93**:917-934

Gilbert, W. (1986) The RNA world, *Nature* **319**:618

Gilmore, T.D. (1991) Malignant transformation by mutant Rel proteins, *Trends Genet.* **7**:318-322

Glisin, V., Crkvenjakov, R. and Byus, C. (1974) Ribonucleic acid isolated by cesium chloride centrifugation, *Biochemistry* **13**:2633-2637

Glover, D.M. (1991) Mitosis in the *Drosophila* embryo- in and out of control, *Trends Genet.* 7:125-131

Goldberg, D.A., Posakony, J.W. and Maniatis, T. (1983) Correct developmental expression of a cloned alcohol dehydrogenase gene transduced into the *Drosophila* germ line, *Cell* **34**:59-73

Gonzales, C., Casal, J. and Ripoli, P. (1988) Functional monopolar spindles caused by mutation in *mgr*, a cell division gene in *Drosophila melanogaster*, *J. Cell Sci.* **89**:34-47

Goodman, C.S., Bastiani, M.J., Doe, C.Q., du Lac, S., Helfand, S.L., Kuwada, J.Y. and Thomas, J.B (1984) Cell recognition during neuronal development, *Science* **225**:1271-1279

Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1988) A conserved NTP-motif in putative helicases, *Nature* **333**:22

Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1989) Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes, *Nucl. Acids Res.* **17**:4713-4730

Gosh, S., Gifford, A.M., Riviere, L.R., Tempst, P., Nolan, G.P. and Baltimore, D. (1990) Cloning of the p50 DNA binding subunit of NF-κB: homology to *rel* and *dorsal*, *Cell* **62**:1019-1029

Graham, A., Papalopulu, N. and Krumlauf, R. (1989) The murine and *Drosophila* homeobox gene complexes have common features of organization and expression, *Cell* **57**:367-378

Green, M.M., Yamamoto, M. and Miklos G.L.G. (1987) Genetic instability in *Drosophila melanogaster*: cytogenetic analysis of MR-induced X chromosome deficiencies, *Proc. Natl. Acad. Sci. USA* **84**:4533-4537

Hall, J.C. and Kankel, D.R. (1976) Genetics of acetylcholinesterase in *Drosophila melanogaster*, *Genetics* **83**:517-535

Hall, L.M.C., Mason, P.J. and Spierer, P. (1983) Transcripts, bands and genes in 315,000 base pairs of *Drosophila* DNA, *J. Mol. Biol.* **169**:83-96

Hanahan, D. (1983) Studies on transformation of *Escherichia coli* with plasmids, *J. Mol. Biol.*, **166**: 557

Hay, E.D. (1984) Collagen and embryonic development, In *Cell Biology of Extracellular Matrix*, (ed. Hay) pp 379-409 New York Plenum Press

Hay, B., Jan, L.Y. and Jan, Y.N. (1988) A protein component of *Drosophila* polar granules is encoded by *vasa* and has extensive sequence similarity to ATP-dependent helicases, *Cell* 55:577-587

Haynes, S.R., Rebert, M.L., Mozer, B.A., Forquignon, F. and Dawid, I.B. (1987) *pen* repeat sequences are GGN clusters and encode a glycine-rich domian in a *Drosophila* cDNA homologous to the rat helix destabilizing protein, *Proc. Natl. Acad. Sci. USA* **84**:1819-1823

Henkemeyer, M.J. Gertler, F.B., Goodman, W. and Hoffmann, F.M. (1987) The *Drosophila Abelson* proto-oncogene homolog: Identification of mutant alleles that have pleiotropic effects late in development, *Cell* **51**:821-828

Herr, W., Sturm, R.A., Clerc, R.G., Corcoran, L.M., Baltimore, D., Sharp, P.A., Ingraham, H.A., Rosenfeld, M.G., Finney, M., Ruvkun, G. and Horvitz, H.R. (1988) The POU domain: a large conserved region in the mammalian *pit*-1, *oct*-1, *oct*-2, and *Caenorhabditis elegans unc*-86 gene products, *Genes Dev* **2**:1513-1516

Highberger, J.H., Corbett, C., Dixit, S.N., Yu, W., Seyer, J.M., Kang, A.H. and Gross, J. (1982) Amino acid sequence of chick skin collagen α1(I)-CB8 and the complete primary structure of the helical portion of the chick skin collagen α1(I) chain, *Biochemistry* **21**:2048-2055

Hilliker, A.J., Clark, S.H., Chovnick, A. and Gelbhart, W.M. (1980) Cytogenetic analysis of the chromosomal region immediately adjacent to the *rosy* locus in *Drosophila melanogaster*, *Genetics* **95**:95-110

Hirling, H., Scheffner, M., Restle, T. and Stahl, H. (1989) RNA helicase activity associated with human p68 protein, *Nature* **339**:562-564

Hoch, M., Schroder, C., Seifert, E. and Jackle, H. (1990) Cis-acting control elements for Kruppel expression in the *Drosophila* embryo, *EMBO J.* **9**:2587-2595

Hodgman, T.C. (1988) A new superfamily of replicative proteins, *Nature* **333**:22-23

Hultmark, D. Klemenz, R. and Gehring, W.J. (1986) Translational and transcriptional control elements in the untranslated leader of the heat-shock gene *hsp22*, *Cell* **44**:429-438

Iggo, R., Picksley, S., Southgate, J., McPheat, J. and Lane, D.P. (1990) Identification of a putative RNA helicase in *E.coli*, *Nucl. Acids Res.* **18**:5413-5417

Ingham, P.W. (1988) The molecular genetics of embryonic pattern formation in *Drosophila*, *Nature* **335**:25-34

Ingham, P.W., Howard, K. R., and Ish-Horowicz, D. (1985) Transcription pattern of the *Drosophila* segmentation gene *hairy*, *Nature* **318**:439-445

John, B. and Miklos, G.L.G. (1988) The eukaryotic genome in development and evolution. *Allen and Unwin*, London

Jeffreys, A.J. and Harris, S. (1984) Processed Pseudogenes *Bioessays* **1**:253-258

Kelly, L.E. (1983) An altered electroretinogram transient associated with an unusual jump response in a mutant of *Drosophila*, *Cell. Mol. Neurobiol.* **3**:143-149

Kessel, M. and Gruss, P. (1990) Murine developmental control genes, *Science* **249**:374-379

Kidd, S., Lockett, T.J. and Young, M.W. (1983) The *Notch* locus of *Drosophila melanogaster*, *Cell* **34**:421-433

Kiernan, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., LeBail, O., Urban, M.B., Kourilsky, P., Baeuerle, P.A. and Israel, A. (1990) The DNA binding subunit of NF-κB is identical to factor KBF I and homologous to the *rel* oncogene product, *Cell* **62**:1007-1018

King, D.S. and Beggs, J.D. (1990) Interactions of PRP2 protein with pre-mRNA splicing complexes in *Saccharomyces cerevisiae*, *Nucl. Acids Res.* **18**:6559-6564

Klemenz, R., Weber, U. and Gehring, W.J. (1987) The *white* gene as a marker in a new P-element vector for gene transfer in *Drosophila*, *Nucl. Acids Res.* **15**:3947-3959

Knibiehler, B., Mirre, C., Cecchini, J. and Le Parco, Y. (1987) Haemocytes accumulate collagen transcripts curing *Drosophila melanogaster* metamorphosis, *Roux's Arch. Dev. Biol.* **196**:243-247

Kozak, M. (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes, *Nucl. Acids Res.* 9:5233-5252

Kramer, J.M., Cox, G.N., and Hirsch, D. (1982) Comparisons of the complete sequences of two collagen genes from *Caenorhabditis elegans*, *Cell* **30**:599-606

Kramers, P.G.N., Schalet, A.P., Paradi, E. and Huiser-Hoogteyling, L. (1983) High proportion of multi-locus deletions among hycanthone induced X linked recessive lethals in *Drosophila melanogaster*, *Mutat. Res.* **107**:187-201

Kuroda, M.I., Kernan, M.J., Kreber, R., Ganetzky, B. and Baker, B.S. (1991) The *maleless* protein associates with the X chromosome to regulate dosage compensation in *Drosophila*, *Cell* **66**:935-947

Lal, A.A., de la Cruz, V.F., Welsh, J.A., Charoenvit, Y., Maloy, W.L. and McCutchan, T.F. (1987) Structure of the gene encoding the circusporozoite protein of *Plasmodium yoelli*, *J. Biol. Chem.* **262**:2937-2940

Lane, D. (1988) Enlarged family of putative helicases, *Nature* **334**:478

Laski, F.A., Rio, D.C. and Rubin, G.M. (1986) Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing, *Cell* **44**:7-19

Laski, F.A. and Rubin, G.M. (1989) Analysis of the cis-acting requirements for germ-line-specific splicing of the P-element PRF2-ORF3 intron, *Genes Dev.* **3**:720-728

Lasko, P.F. and Ashburner, M. (1988) The product of the *Drosophila* gene *vasa* is very similar to eukaryotic initiation factor-4A, *Nature* 335:611-617

Lefevre, G. (1974) The relationship between genes and polytene chromosome bands, *Ann. Rev. Genet.* **8**:51-62

Lefevre, G. (1976) The polytene chromosomes, In *The genetics and biology of Drosophila melanogaster*, Vol1A, (eds. Ashburner, M. and Novitski, E.) Academic Press, New York, pp31-66

Lefevre, G. (1981) The distribution of randomly recovered X-ray-induced sex-linked genetic effects in *Drosophila melanogaster*, *Genetics* **99**:461-480

Lefevre, G. and Watkins, W. (1986) The question of the total gene number in *Drosophila melanogaster*, *Genetics* **113**:868-895

Le Parco, Y., Cecchini, J., Knibiehler, B. and Mirre, C. (1986) Characterization and expression of collagen-like genes in *Drosophila melanogaster*, *Biol. Cell* **56**:217-226

Leroy, P., Alzari, P., Sassoon, D., Wolgemuth, D. and Fellous, M. (1989) The protein encoded by a murine male germ cell-specific transcript is a putative ATP-dependent RNA helicase, *Cell* **57**:549-559

Lewis, R.A., Kaufman, T.C., Denell, R.E. and Tallerico, P. (1980) Genetic analysis of the *Antennapedia* gene complex (*ANT-C*) and adjacent chromosomal regions of *Drosophila melanogaster*, I. Polytene chromosome segments 84B-D, *Genetics* **95**:367-381

Lifschytz, E. and Falk, R. (1968) Fine structure analysis of a chromosome segment in *Drosophila melanogaster*, Analysis of X-ray induced lethals, *Mutat. Res.* **6**:235-244

Lifschytz, E. and Falk, R. (1969) Fine structure analysis of a chromosome segment in *Drosophila melanogaster*, Analysis of ethyl methanesulfonate- induced lethals, *Mutat. Res.* **8**:147-155

Lifschytz, E. and Yakobovitz, N. (1978) The role of X-linked lethal and viable sterile mutations in male gametogenesis of *Drosophila melanogaster*: genetic analyses, *Mol. Gen. Genet.* **161**:275-284

Linder, P. and Slonimski, P. (1989) An essential yeast protein, encoded by duplicated genes *TIF1* and *TIF2* and homologous to the mammalian translation initiation factor eIF-4A, can suppress a mitochondrial missense mutation, *Proc. Natl. Acad. Sci. USA* **86**:2286-2290

Linder, P., Lasko, P.F., Ashburner, M., Leroy, P., Nielsen, P.J., Nishi, K., Schnier, J. and Slonimski, P. (1989) Birth of the D-E-A-D box, *Nature* **337**:121-122

Luhbahn, D.B., Joseph, D.R., Sullivan, P.M., Huntington, F.W., French, F.S. and Wilson, E.M. (1988) Cloning of human androgen receptor complementary DNA and localization to the X chromosome, *Science* **240**:327-330

Maniatis, T., Fritsch, E. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, New York, Cold Spring Harbor Laboratory Press

Manley, J.L. (1988) Polyadenylation of mRNA precursors, *Biochim. Biophys. Acta* **950**:1-12

Markov, T.A. and Merriam, J. (1977) Phototactic and geotactic behaviour of countercurrent defective mutants of *Drosophila melanogaster*, *Behav. Genet.* **7**:447-455

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA, *Proc. Natl. Acad. Sci. USA* **74**:560-564

McMaster,G.K. and Carmichael, G.G. (1977) Analysis of single- and doublestranded nucleic acids on polyacrylamide and agarose gels by using glyoxal and acridine orange, *Proc. Natl. Acad. Sci.* **74**:4835

McGinnis, W., Garber, R.L., M., Wirz, J., Kuroiwa, A. and Gehring, W.J. (1984a) A homologous protein coding sequence in *Drosophila* homeotic genes and its conservation in the metazoans, *Cell* **37**:403-408

McGinnis, W., Levine, M.S., Hafen, E., Kuroiwa, A. and Gehring, W.J. (1984b) A conserved DNA sequence in homeotic genes of *Drosophila Antennapedia* and *bithorax* complexes, *Nature* **308**:428-433

Miklos, G.L.G., Kramers, P.G.N. and Schalet, A.P. (1986) The proximal-distal orientation of two lethal complementation groups *A112* and *LB20* in region 19F at the base of the X chromosome, *Dros. Inf. Serv.* **63**:96-97

Miklos, G.L.G., Kelly, L.E., Coombe, P.E., Leeds, C. and Lefevre G. (1987) Localization of the genes *shaking-B, small optic lobes, sluggish-A, stoned* and *stress-sensitive-C* to a well-defined region on the X chromosome of *Drosophila melanogaster*, *J. Neurogenet.* **4**:1-19

Mirre, C., Cecchini, J., Le Parco, Y. and Knibiehler, B. (1988) De novo expression of a type IV collagen gene in *Drosophila* embryos is restricted to mesodermal derivatives and occurs at germ band shortening, *Development* **102**:369-376

Mitchell, P.J. and Tijan, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins, *Science* **245**:371-378

Mount, S.M. (1982) A catalogue of splice sequences, *Nucl. Acids Res.* **10**:459-472

Mozer, B., Marlor, R., Parkhurst, S. and Corcef W. (1985) Characterisation and developmental expression of a *Drosophila ras* oncogene, *Mol. Cell Biol.* **5**:885-889

Muller, H.J. (1927) Artificial transmutation of the gene, *Science* **66**:84-87

Mullins, M., Rio, D.C. and Rubin, G. (1989) cis-acting DNA sequence requirements for P-element transposition, *Genes Dev.* **3**:729-738

Natzle, J.E., Monson, J.M. and McCarthy, B.J. (1982) Cytogenetic location and expression of collagen-like genes in *Drosophila*, *Nature* **296**:368-371

Nielsen, P.J. and Trachsel, H. (1988) The mouse protein synthesis initiation factor 4A gene family includes two related functional genes which are differentially expressed, *EMBO J.* **7**:2097-2105

Nielsen, P.J., McMaster, G.K. and Trachsel. H. (1985) Cloning of eukaryotic protein synthesis initiation factor genes: isolation and characterization of cDNA clones encoding factor eIF-4A, *Nucl. Acids Res.* **13**:6867-6881

Nishi, K., Morel-Deville, F., Hershey, J.W.B., Leighton, T. and Schnier, J. (1988) An eIF-4A like protein is a suppressor of an *Escherichia coli* mutant defective in 50S ribosomal subunit assembly, *Nature* **336**:496-498

Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in *Drosophila*, *Nature* **287**:795-801

O'Hare, K. and Rubin, G.M. (1983) Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome, *Cell* **34**:25-36

O'Hare, H., Levis, R. and Rubin, G.M. (1983) Transcription of the white locus in *Drosophila melanogaster*, *Proc. Natl. Acad. Sci. USA* **80**:6917-6921

Olsen, B.R. (1981) Collagen Biosynthesis. In *Cell Biology of the extracellular matrix* (E.D. Hay, ed) pp. 139-177. New York: Plenum,

Painter, T.S. (1933) A new method for the study of chromosome rearrangements and the plotting of chromosome maps, *Science* **78**:585-586

Pankratz, M.J. and Jaeckle, H. (1990) Making stripes in the *Drosophila* embryo, *Trends Genet.* 6:287-292

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA* **85**:2444-2448

Perlman, D. and Halvorson, H.O. (1983) A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides, *J. Mol. Biol.* **167**:391-409

Perrimon, N., Engstrom, L. and Mahowald, A.P. (1984) Developmental genetics of the 2E-F region of the *Drosophila* X chromosome: a region rich in "developmentally important" genes, *Genetics* **108**:559-572,

Perrimon, N., Engstrom, L. and Mahowald, A.P. (1985) Developmental genetics of the 2C-D region of the *Drosophila* X chromosome, *Genetics* **111**: 23-41.

Perrimon, N., Smouse, D. and Miklos, G.L.G. (1989) Developmental genetics of the base of the X chromosome of *Drosophila melanogaster* *Genetics* **121**:313-331

Pfeifer, M., Karch, F. and Bender, W. (1987) The *bithorax* complex: control of segmental identity, *Genes and Dev* **1**:891-898

Pirrotta, V. (1986) Cloning *Drosophila* genes, In *Drosophila a practical approach* (ed Roberts, D.B.), IRL Press, pp 83-109

Politz, J.C. and Edgar, R.S. (1984) Overlapping stage specific sets of numerous small collagenous polypeptides are translated *in vitro* from *Caenorhabditis elegans* RNA, *Cell* **37**:853-860

Poole, S.J., Kauvar, L.M., Drees, B. and Kornberg, T. (1985) The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. *Cell* **40**:37-43

Proudfoot, N.J. (1991) Poly(A)signals, *Cell* **64**:671-674

Proudfoot, N.J. and Brownlee, G.G. (1976) 3' non-coding region sequences in eukaryotic mRNA, *Nature* **263**:211-214

Pyati, J. (1976) Cytological localization of the maternal effect gene *tuh-1* in *Drosophila melanogaster*, *Mol.Gen.Genet.* **146**:189-190

Ramachandran, G.N. (1967) Structure of collagen at the molecular level, In *treatise on collagen*, 1 (ed. Ramachandran) New York, Academic Press pp 103-183

Rautenberg, J., Timpl, R. and Furthmayr, H. (1972) Structural characterization of N-terminal antigenic determinants in calf and human collagen, *Eur. J. Biochem.* **27**:231-237

Rijsewijk F., Schuermann M., Wagenaar E., Parren P., Weigel D. and Nusse R. (1987) The *Drosophila* homolog of the mouse mammary oncogene *int*-1 is identical to the segment polarity gene *wingless*, *Cell* **50**:649-657

Rio, D.C. (1991) Regulation of *Drosophila* P element transposition, *Trends Genet.* **7**:282-287

Ripoli, P., Pimpinelli, S., Valdivia, M.M. and Avila, J. (1985) A cell division mutant of *Drosophila* with a functionally abnormal spindle, *Cell* **41**:907-912

Roberts, D.B. (1986) Basic *Drosophila* care and techniques, In *Drosophila a practical approach* (ed Roberts, D.B.), IRL Press, pp1-39

Royden, C.S., Pirrotta, V. and Jan, L.Y. (1987) The *tko* locus, site of a behavioral mutation in *D. melanogaster*, codes for a protein homologous to prokaryotic ribosomal protein S12, *Cell* **51**:165-173

Ruben, S.M., Dillon, P.J., Schreck, R., Henkel, T., Chen, C., Mahler, M., Baeuerle, P.A. and Rosen, C.A. (1991) Isolation of a *rel*-related human cDNA that potentially encodes the 65-kD subunit of NF-κB, *Science* **251**:1490-1493

Rubin, G.M. and Spradling, A.C. (1982) Genetic transformation of *Drosophila* with transposable element vectors, *Science* **218**:348-353

Sambrook, J., Fritsch, E.F. and Maniatis, T. (eds.) (1989): *Molecular Cloning: a laboratory manual*; second edition, Cold Spring Harbor Laboratory Press

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain termination inhibitors, *Proc. Natl. Acad. Sci. USA* **74**:5463-5476

Schalet, A.P. (1986) The distribution of and complementation between spontaneous X-linked recessive lethal mutations recovered from crossing long term laboratory stocks of *Drosophila melanogaster*, *Mutat. Res.* **163**:115-144

Schalet, A. and Finnerty, V. (1968) New mutants: Report of A.Schalet and V.Finnerty. *Dros. Inf. Serv.* **43**:65-66

Schalet, A. and Lefevre, G. (1973) The localization of "ordinary" sex-linked genes in section 20 of the polytene X chromosome of *Drosophila melanogaster*, *Chromosoma* **44**:183-202

Schalet, A. and Lefevre, G. (1976) The proximal region of the X-chromosome, In *The genetics and biology of Drosophila melanogaster*, Vol1B, (eds. Ashburner, M. and Novitski, E.) Academic Press, New York, pp848-902

Schalet, A. and Singer, K. (1971) A revised map of genes in the proximal region of the X chromosome of *Drosophila melanogaster*. *Dros. Inf. Serv.* **46**:131-132

Schwer, B. and Guthrie, C. (1991) PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome, *Nature* **349**:494-499

Scott, M.P. and Weiner, A.J. (1984) Structural relationships among genes that control development: Sequence homology between the *Antennapedia*, *Ultrabithorax* and *fushi tarazu* loci of Drosophila, *Proc. Natl. Acad. Sci.* **81**:4115-4119

Seraphin, B., Simon, M., Boulet, A. and Faye, G. (1989) Mitochondrial splicing requires a protein from a novel helicase family, *Nature* **337**:84-87

Serfling, E., Jasmin, M. and Schaffner, W. (1985) Enhancers and eukaryotic gene transcription, *Trends Genet.* **1**:224-230

Seyer, J.M. and Kang, A.H. (1977) Covalent structure of collagen: amino acid sequence of cyanogen bromide peptides from the amino-terminal segment of type III collagen of human liver, *Biochemistry* **16**:1158-1164

Shannon, M.P., Kaufman, T.C., Shen, W.M. and Judd, B.H. (1974) Lethality patterns and morphology of selected lethals and semi-lethal mutations in the *zeste-white* region of *Drosophila melanogaster Genetics* **72**:615-638.

Shaw, D.R., Richter, H., Giorda, R., Ohmachi, T. and Ennis, H.L. (1989) Nucleotide sequences of *Dictyostelium discoideum* developmentally regulated cDNAs rich in (AAC) imply proteins that contain clusters of asparagine, glutamine or threonine, *Mol. Gen. Genet* **218**:453-459

Shearn, A., Rice, T., Garen, A. and Gehring, W. (1971) Imaginal disc abnormalities in lethal mutants of *Drosophila*, *Proc. Natl. Acad. Sci.* **71**:1393-1397

Siebel, C.W. and Rio, D.C. (1990) Regulated splicing of the *Drosophila* P transposable element third intron *in vitro*: somatic repression, *Science* **248**:1200-1208

Smith, D.A., Baker, B.S. and Gatti, M. (1985) Mutations in genes encoding essential mitotic functions in *Drosophila melanogaster*, *Genetics* **110**:647-670

Snyder, M. (1989) The SPA2 protein of yeast localizes to sites of cell growth, *J. Cell Biol.* **108**:1419-1429

Southern, E. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, *J. Mol. Biol.* **98**:503-517

Spierer, P., Spierer, A., Bender, W. and Hogness, D.S. (1983) Molecular mapping of genetic and chromomeric units in *Drosophila melanogaster*, *J. Mol. Biol.* **168**:35-50.

Spradling A.C. (1986) P element-mediated transformation, In Drosophila A practical approach (ed. Roberts) IRL press, pp175-199

Spradling, A.C. and Rubin, G.M. (1981) *Drosophila* genome organization: conserved and dynamic aspects, *Ann. Rev. Genet* **15**:219-264

Spradling, A.C. and Rubin, G.M. (1982) Transposition of cloned P elements into *Drosophila* germ line chromosomes, *Science* **218**:341-347

Spradling, A.C. and Rubin, G.M. (1983) The effect of chromosomal position on the expression of the *Drosophila* xanthine-dehydrogenase gene, Cell **34**:47-57

Stacey, A., Bateman, J., Choi, T., Mascara, T., Cole, W. and Jaenisch, R. (1988) Perinatal lethal osteogenesis imperfecta on transgenic mice bearing an egineered mutant pro-α1(I)collagen gene, *Nature* **332**:131-136

Staden, R. (1984) Computer methods to locate signals in nucleic acids, *Nucl. Acids Res.* **12**:505-520

Steward, R. (1987) *Dorsal*, an embryonic polarity gene in *Drosophila*, is homologous to the vertebrate proto-oncogene, c-*rel*, *Science* **238**:692-695

Steward, R. and Nusslein-Volhard, C. (1986) The genetics of the *dorsal-Bicaudal-D* region of *Drosophila melanogaster*, *Genetics* **113**:665-678

Struhl, K. (1987) Promoters, activator proteins and the mechanism of transcriptional activation in yeast, *Cell* **49**:295-297

Sunkel, C. and Glover, D.M. (1988) *polo*, a mitotic mutant of *Drosophila* displaying abnormal spindle poles, *J Cell Sci.* **89**:25-38

Tearle, R.G. (1987) Genetics and biochemistry of ommochrome biosynthesis in *Drosophila melanogaster*, PhD thesis, The Australian National University, Canberra

Thisse, C., Perrin-Schmitt, F., Stoetzel, C. and Thisse, B. (1991) Sequence-specific transactivation of the *Drosophila* twist gene by the dorsal gene product. *Cell* **65**:1191-1201

Thomas, P.S. (1983) Hybridization of denatured RNA transferred or dotted to nitrocellulose paper, *Methods Enzymol.* **100**:255

Thomas, J.B. and Wyman, R.J. (1984) Mutations altering synaptic connectivity between identified neurons in *Drosophila*, *J. Neurosci.* **4**:350-538

Thomas, J.B., Bastiani, M.J., Bate, M. and Goodman, C.S. (1984) From grasshopper to *Drosophila*: a common plan for neuronal development, *Nature* **310**:203-207

Thomas, J.B., Crews, S.T. and Goodman, C.S. (1988) Molecular genetics of the *single-minded* locus: a gene involved in the development of the *Drosophila* nervous system, *Cell* **52**:133-141

Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W.J. and Goodman, H.M. (1977) Rat insulin genes: Construction of plasmids containing the coding sequences, *Science* **196**:1313-1319

Van Loon, A.P.G.M., DeGroot, R.J., DeHaan, M., Dekker, A. and Grivell, L.A. (1984) The DNA sequence of the nuclear gene coding for the 17 kd subunit of the yeast ubiquinol cytochrome c reductase: a protein with an extremely high content of acidic amino acids, *Embo J.* **3**:1039-1043

Vanin, E.F. (1984) Processed pseudogenes; characteristics and evolution, *Biochimica et Biophysica Acta*, **782**:231-241

Von Heijne, G. (1985) Signal sequences: the limits of variation, *J. Mol. Biol.* **184**:99-105

Wagner, M. (1986) A consideration of the origin of processed pseudogenes, *Trends Genet.* **2**:134-137

Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.* **1**:945-951

Wharton, K.A., Yedvobnick, B., Finnerty, V.G. and Artavanis-Tsakonas, S. (1985) *opa*: A novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in *D. melanogaster*, *Cell* **40**:55-62

Whiting, J.H., Pliley, M.D., Farmer, J.L. and Jeffery, D.E. (1989) *In situ* hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis*, *Genetics* **122**:99-109

Wickens, M. and Stephenson, P. (1984) Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3'end formation, *Science* **226**:1045-1051

Wiedmann, M., Kurzchalia, T.V., Hartmann, E. and Rapoport, T.A. (1987)A signal sequence receptor in the endoplasmatic reticulum membrane, *Nature* **328**:830-833

Wieschaus, E., Audit, C. and Masson, M. (1981): A clonal analysis of the roles of somatic cells and germ line during oogenesis in *Drosophila*, *Dev. Biol.* **88**:92-103

Woolf, C.M. (1968) Male genital disc defect in *Drosophila melanogaster*, *Genetics* **60**:111-121

Zusman, S.B. and Wieschaus, E. (1985) Requirements for zygotic gene activity during gastrulation in *Drosophila melanogaster*, *Dev. Biol.* **111**:359-371